Univerzita Karlova v Praze

Filozofická fakulta

Ústav informačních studií a knihovnictví

Informační věda – Informační studia a knihovnictví

Jan H u t a ř

Digitalizace, popis pomocí metadat a jejich formáty

Digitization, metadata description and metadata formats

Disertační práce

Vedoucí práce – Stanislav Kalkus, Ph.D.

2012

P r o h l á š e n í

Prohlašuji, že jsem dizertační práci vypracoval samostatně, že jsem řádně citoval všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

V Praze, 23. března 2012 ……

podpis

„Digital information lasts forever, or five years – whichever comes first"

Jeff Rothenberg

Abstrakt (CZ)

Disertační práce je věnována problematice digitalizace, metadatového popisu a souvislostem, které tyto procesy spojují. V posledních letech se k těmto běžným tématům pojí také logická dlouhodobá ochrana digitálních dat, která je na metadatech a tedy i na procesech digitalizace, kde metadata vznikají, do velké míry závislá. První úvodní kapitoly disertační práce se zabývají digitalizací, teoretickými a praktickými problémy dlouhodobé ochrany digitálních dat s důrazem na metadata. Rozebrán je z pohledu metadat i referenční rámec OAIS, který je východiskem pro dnešní podobu ochranných metadat i podobu digitálních repozitářů. Metadatům se věnují také další kapitoly disertační práce. Je analyzován obecný vývoj metadat s důrazem na administrativní, ochranná a technická metadata používaná v paměťových institucích. Podobné hledisko má i následující kapitola o využívání metadat v Národní knihovně České republiky (dále NK ČR). Ta popisuje vývoj používání metadat ve dvou obdobích až do současnosti a nabízí i komentáře k tomu, jak praxe a používání standardů metadat reflektovaly potřeby dlouhodobé ochrany digitálních dat. Jedna z posledních částí práce se zabývá způsoby uložení dat a problematikou certifikace repozitářů. Praktickou částí práce a výstupem několikaletého výzkumu je návrh metadatového profilu pro masovou (robotickou) digitalizaci novodobých monografií a periodik v projektu *Národní digitální knihovna* (NDK). Profil obsahuje schémata METS, PREMIS, Dublin Core, MIX, MODS a ALTO XML. Návrh byl v méně rozpracované podobě součástí Zadávací dokumentace projektu NDK v létě 2011. Od podzimu 2011 se podle profilu vytvářejí také metadata v projektu NK ČR ANL+ (analytické zpracování periodik).

Klíčová slova:
digitalizace; dlouhodobá ochrana digitálních dat; metadata; aplikační profil metadat; důvěryhodný digitální repozitář; nástroje; OAIS; PREMIS; MIX; METS; MODS; ALTO XML

Abstract (EN)

This thesis is dedicated to the processes of digitization and metadata description, as well as the links that connect them. In recent years, another topic has been becoming relevant for both mentioned processes – the logical long-term preservation of digital objects. Long-term preservation is dependent on metadata and therefore on the processes of digitization, when some important metadata is created.

The first introductory chapter of the thesis briefly describes, with an emphasis on metadata, digitization and theoretical and practical problems of long-term preservation of digital objects. The OAIS reference framework is also analysed, since it is the background for digital preservation and present preservation metadata standards. OAIS is also important for the shape and functionality of digital repositories. Metadata is also the topic of the next chapter of the thesis. General metadata use and its development are discussed, with emphasis on administrative, technical and preservation metadata. The following chapter focuses on the use of metadata in the National Library of the Czech Republic. It describes the evolution during two periods leading up to the present. This section includes comments on how long-term preservation has been reflected in used metadata standards. The second-to-last part of the thesis deals with data storage possibilities and the issue of digital repository certification. The practical part of the work and the research output is a metadata application profile proposal for mass (robotic) digitization of modern monographs and periodicals for the National Digital Library project (NDK). The profile

contains metadata schemas METS, PREMIS, Dublin Core, MIX, MODS and ALTO XML. The proposal was (in a less developed form) part of the call for tender documentation for NDK projects in summer 2011. Since autumn 2011, metadata from the project ANL+ (periodicals analytical description) has been created according to that profile.

Keywords:
digitisation; long-term preservation; metadata; metadata application profile; trusted digital repository; tools; OAIS; PREMIS; MIX; METS; MODS; ALTO XML

1. Cíle a předmět práce

Myslím, že dnes již není pochyb o tom, že dnešní společnost již není společností „papírovou“, nýbrž elektronickou. Velká část vědění, která byla tradičně uložena na papírových nosičích, je nyní přístupná pouze ve formě elektronické (digitální). Tato skutečnost se zdá býti výhodou, hlavně uživatelsky, ovšem představuje také velkou hrozbu. Hrozbu nenávratné ztráty informací, vědění, zkušeností, poznatků. V posledních letech si hlavně pracovníci tzv. paměťových institucí začali uvědomovat, že toto nebezpečí je reálné a že ztráty digitálních informací jsou současným problémem nejen jejich institucí, ale každého z nás. Někteří z nich pochopili, že bude potřeba od základu změnit způsob pohledu na paměťové instituce a dokumenty v nich uchovávané, který se naprosto liší od klasického pohledu, takříkajíc statického, kdy vše bylo dostupné v papírové kopii a hlavní starostí knihovny bylo shromáždit určitá díla a uchovat je zajištěním odpovídajícího prostředí uložení. To se s příchodem digitálních dokumentů/objektů změnilo. Předkládaná disertační práce popisuje mj. tento společenský a technologický jev a to, jak se s ním vyrovnávají paměťové instituce, především knihovny ve světě i u nás.

Cílem disertační práce je popsat současný stav a souvislosti procesů digitalizace analogových předloh tištěných dokumentů, tvorby metadat a dlouhodobé ochrany takto vzniklých digitálních objektů. Je popsána důležitost metadat a jejich specifikace pro proces digitalizace, ale i pro aktivity dlouhodobého uložení, pro které jsou metadata pojítkem s digitalizací. Údaje určující pro dlouhodobou ochranu totiž vznikají již při tvorbě digitálního objektu. Práce se snaží poskytnout náhled na problematiku dlouhodobé ochrany digitálních dat (digital preservation) ve spojitosti s metadaty, která během digitalizace vznikají a s jejich dalším „životem“ v digitálním archivu. Mnoho let byla situace taková, že se digitalizovaly dokumenty, vytvářela se popisná metadata a nemyslelo se na problémy, které přinese skutečnost, že digitální objekty podléhají zastarávání SW i HW. Nevznikala ani technická, ani administrativní metadata. Toto se autor během svého působení v NK ČR snažil změnit (viz návrhy použití PREMIS a MIX). V tomto smyslu se tedy práce věnuje problematice *digital curation*, která se dle Harveyho [HARVEY, 2010] zabývá celým životním cyklem digitálního objektu, včetně plánování metadat před jeho vznikem tak, aby bylo možno zajistit správu, aktivní (rozuměj logickou) ochranu a zpřístupnění digitálních objektů. Dlouhodobá ochrana digitálních dat (*digital preservation*), je podskupinou *digital curation*, a jejím cílem je dlouhodobé uchování digitálních dokumentů v autentické a použitelné podobě pro budoucí uživatele. Práce analyzuje procesy, které se s metadaty dějí v digitálním úložišti v LTP (*Long-Term Preservation*) systému.

Výsledkem dlouholeté práce a hlavní částí této disertační práce je návrh nové specifikace metadat (metadatového profilu) pro projekt *Národní digitální knihovna* (NDK). Jde o specifikaci pro digitalizaci monografií a periodik, založenou na standardech METS, PREMIS, MIX, ALTO XML a MODS. Tedy na standardech, které se dnes běžně ve světě používají a pokrývají všechny aspekty metadat potřebných pro bezpečné uložení, zpřístupnění, správu a především ochranu v dlouhodobém horizontu desítek a možná i stovek let od okamžiku digitalizace.

První část textu práce obsahuje obecné části o metadatech, dlouhodobé ochraně digitálních dat a důvěryhodných úložištích. Tyto oblasti spolu souvisejí navzájem. Jsou popsány jednotlivé relevantní standardy a jejich možnosti s důrazem na ty, které jsou součástí metadatové

specifikace pro projekt NDK. Naopak popis jednotlivých standardů metadat, které nejsou součástí specifikace, je schematický (např. v kapitole o vývoji metadat) a nedělá si nároky na úplnost.[1] Tato část práce je přehledová a jsou popsány pouze nejvýznamnější zástupy metadatových standardů používaných v oblasti paměťových institucí; je založena na analýzách publikovaných textů a také na zkušenostech autora.
Druhá část disertační práce mapuje vývoj používání metadat v NK ČR a vývoj chápání a opatření vedoucích k současnému stavu, kdy se NK ČR snaží svá data opatřit všemi typy metadat tak, aby byla použitelná i v budoucnu. To je nutné považovat za veliký pokrok od „pouhé" digitalizace. U schémat, která byla nebo jsou historicky používána v projektech NK ČR a u schémat, se kterými se počítá pro využití v LTP systému, je podrobnější popis a analýza. Druhá část se také věnuje způsobu uložení dat a tomu, jakou roli hrají při uložení metadata. Tato problematika je popsána spolu s obecným vývojem, vývojem pohledu na uložení dat v NK ČR. Je také rozebrána problematika tzv. certifikovaných důvěryhodných repozitářů. Text této části vznikal analytickým rozborem metadatových specifikací, výzkumných projektů a existujících aplikací.
Třetí a hlavní částí disertační práce jsou pak samotné metadatové profily pro monografie a periodika. Tato část je hlavním přínosem celého textu a výzkumu. Profily byly vytvářeny v letech 2009-2011 v několika etapách, od prostého návrhu až po podrobnou specifikaci, která je předkládána. V roce 2011 se podle profilu pro periodika začalo digitalizovat v projektu ANL+, který digitalizuje a analytickou formou zpracovává periodické publikace. Jde o několikaletý projekt, jehož výstupy jsou mj. dostupné i v Jednotné informační bráně[2] (JIB). Metadatový profil je prezentován v textové podobě s vysvětlením jednotlivých elementů a celé koncepce. Nebylo cílem práce popsat technické řešení vlastní tvorby metadat podle profilu.

Metody použité při psaní práce jsou vlastní výzkum a návrh specifikace, studium podobných řešení a implementací ve světě, studium publikovaných pramenů.

Autor se snažil v disertační práci nezabíhat do technických detailů HW a SW. Jde spíše o návodný text, který by měl dát představu o novém vývoji v oblasti metadat pro digitální objekty, která mají umožnit jejich uložení, správu a dlouhodobou ochranu pro budoucí uživatele. Z tohoto důvodu se práce nevěnuje více metadatům pro zpřístupnění, důraz je naopak kladen na metadata pro archivaci dokumentů vznikajících digitalizací analogové předlohy (monografie, periodika), která jsou v dalších krocích životního cyklu doplňována v archivu, případně LTP systému. Také proto nejsou blíže popsány technologie ani koncepční podstata digitálních knihoven, jejich podrobný vývoj a obrazové formáty v nich používané. V textu práce jsou zmíněny konkrétní nástroje na tvorbu metadat, jejich bližší popis ani analýza ovšem nejsou cílem této práce. Vědomě nejsou řešena metadata pro archiválie, zvukové a video dokumenty, ani pro tzv. digital-born dokumenty, kterými jsou např. uložené webové stránky, kvalifikační práce v digitálních repozitářích univerzit.

Názvy metadatových elementů v textu jsou uvedeny vždy s počátečním velkým písmenem, případně mohou být elementy zapsány v zalomených závorkách, např. <title>. Hodnoty elementů a atributů jsou vždy v uvozovkách. Všechny citace jsou v uvozovkách a kurzívou, pokud jsou přeloženy z originálního znění, učinil tak autor práce. Kurzívou jsou uvedeny také ne zcela běžné

[1] Podstatné informace o vývoji těchto standardů lze najít v mnoha zdrojích na Internetu a v odborných publikacích.
[2] http://info.jib.cz/news/anl

anglické a české termíny, v textu citované zdroje (články, zprávy aj.), názvy projektů a zahraničních organizací (ne ovšem již názvy standardů, SW aplikací nebo systémů). Výrazy české terminologie u popisu referenčního rámce OAIS a standardu PREMIS pro přehlednost začínají velkým písmenem, často jsou doprovázeny pro jednoznačnost i původním anglickým výrazem.
Pro citování použitých informačních zdrojů je v disertační práci použit tzv. Harvardský systém citování, tedy citace pomocí prvního údaje záznamu s datem vydání dokumentu. Pro větší přehlednost jsou použity pro citace v textu hranaté závorky a záhlaví (jména autorů a korporací) jsou uvedena velkými písmeny. Citace, které mají jako záhlaví název zdroje, jsou malými písmeny.

1.1 Odborná základna pro disertační práci – přehled

Podklady pro text předkládané disertační práce vznikaly v průběhu celého doktorandského studia a to převážně díky účasti autora v relevantních projektech v rámci jeho zaměstnání v NK ČR. K tématu dlouhodobé ochrany od roku 2006 publikoval odborné články a prezentace (viz seznam literatury), ze kterých vychází text několika kapitol. Články vznikaly v rámci mezinárodních projektů, i v rámci programů *Výzkumu a vývoje*, které byly řešeny v NK ČR, např. *Budování vzájemně kompatibilních informačních systémů pro přístup k heterogenním informačním zdrojům a jejich zastřešení prostřednictvím Jednotné informační brány* (2004-2010, Ministerstvo kultury ČR), vývoj aplikace zpřístupnění Kramerius od roku 2007 v oblasti specifikace metadat spolu s Knihovnou Akademie věd ČR aj. Velká část publikovaných odborných textů vznikla v souvislosti s plánováním projektu *Národní digitální knihovna*, kterého se autor účastnil od samého počátku (2007) až do konce roku 2011. V projektu NDK byl zodpovědný za vedení pracovní skupiny pro dlouhodobou ochranu digitálních dat a vytvořil také specifikace (metadatové aplikační profily) pro digitalizaci monografií a periodik v rámci projektu. Tyto specifikace jsou nosnou částí předkládané disertační práce. Navazují na předchozí aktivity, kdy byl autor zodpovědný za implementaci standardů METS do aplikace Kramerius (mapování DTD do METS 2007-2008) a ve stejném období vytvořil první specifikaci využití ochranných metadat PREMIS a technických metadat MIX pro digitalizaci novodobých monografií a periodik v NK ČR [HUTAŘ, 2008a] – viz kapitola 5.3.2.
Další pracovní zkušenosti v oblasti dlouhodobé ochrany digitálních dat autor načerpal účastí v projektu *DigitalPreservationEurope*, kterého se NK ČR účastnila v letech 2006-2009 a také jako vedoucí Odboru digitální ochrany tamtéž. Díky zaměstnání v NK ČR se autor každoročně účastnil mezinárodních konferencí k tématu digitálních knihoven a dlouhodobé ochrany digitálních dat (iPRES, Archiving, ECDL/TPDL aj.) a je v dennodenním kontaktu s odborníky z celého světa, se kterými byla většina záležitostí pro projekt NDK konzultována, případně porovnávána se stejnými řešeními.
K tvorbě práci byly využity i zkušenosti získané z návštěv zahraničních knihoven, které jsou v čele ve výzkumu a nasazení řešení dlouhodobé ochrany digitálních dat a také vedení pracovních skupin na úrovni NK ČR, případně na národní úrovni, jako byla *Pracovní skupina PID pro jednoznačné identifikátory*, *Pracovní skupina pro digitální repozitář*, které vznikly v NK ČR a byly složeny z členů z největších českých knihoven.

Účast na projektech EU:

- *Web Cultural Heritage* – v rámci programu *Culture2000*; spolupráce s Estonskou národní knihovnou, Národní a univerzitní knihovnou v Ljublani a Univerzitní knihovnou v Bratislavě; říjen 2005 – říjen 2006

- *DigitalPreservationEurope* – DPE, NK ČR jako jeden z partnerů vedle nejvýznamnějších evropských knihoven; duben 2006 – duben 2009
- *LTP working group* – iniciativa vyvolaná Národní knihovnou Nizozemí; specifikace požadavků na LTP systém použitelný napříč národními knihovny v EU; 2009-2010

Účast na národních projektech výzkumu a vývoje:

- *Budování vzájemně kompatibilních informačních systémů pro přístup k heterogenním informačním zdrojům a jejich zastřešení prostřednictvím Jednotné informační brány* (2004-2010)
- *Optimalizace nástrojů pro digitalizaci tištěných dokumentů ohrožených degradací kyselého papíru* (2006-2010)
- *Ochrana a trvalé zpřístupnění webových zdrojů jako součásti národního kulturního dědictví* (2006-2011)
- projekt *Národní digitální knihovna* (2007-2011)

Přehled odborného působení

2004	ukončení magisterského studia UISK FF UK Praha
	nástup do NK ČR, Oddělení elektronických online zdrojů (WebArchiv)
2005	šestiměsíční stáž na European University Institute, Florencie, Itálie
2007	navržen na cenu konference Inforum za překlad Dublin Core do češtiny
2006 – 2008	vedoucí Referátu pro digitální knihovnu NFS v NK ČR
2008 – 2011	vedoucí Odboru digitální ochrany NK ČR
	člen pracovního týmu plánování projektu NDK, oblast metadat a LTP systému
2007 – 2012	vyučující Ústavu informačních studií a knihovnictví, FF UK, Praha
2009 – 2011	vedoucí pracovní skupiny pro LTP systém projektu NDK
2012 –	Archives New Zealand, Senior Advisor, Digital Continuity Team

2. Úvod

Knihovny byly v minulosti zaměřeny na uchování minulosti, tedy dokumentů starých i současných pro budoucí uživatele. V současné době, s příchodem digitalizace a dlouhodobé ochrany digitálních dat, se knihovny a vlastně všechny paměťové instituce orientují směrem do budoucna. Nepřestaly se starat o dokumenty, ale stávají se vedoucími institucemi ve vývoji technologií, které k tomu potřebují. Jde o jeden z podstatných obratů, který za poslední desetiletí knihovny potkal. Zatímco automatizace v 60. letech 20. století pomohla provádět lépe, levněji a rychleji stávající procesy v knihovnách známé stovky let (katalogizace, akvizice aj.), tak až digitalizace a ochrana digitálních dat přinesla změny těchto procesů.

Instituce všech druhů, mezi nimi i paměťové instituce, vládní úřady, státní správa obecně, produkují a přijímají více a více digitálních dat, která potřebují spravovat, používat a třídit. U digitálních objektů přístupných v digitálních knihovnách uživatelé předpokládají a očekávají, že jednou dostupné dokumenty budou dostupné a použitelné stále, podobně jako knihy, za deset, padesát nebo sto let. Digitální objekty, které tvoří digitální dokumenty, jsou ovšem velmi křehké a ovlivnitelné změnami HW, SW nebo jinými technologickými zvraty. V posledních letech proto probíhá posun od pouhého generování obsahů pomocí digitalizace k přijímání odpovědnosti za dlouhodobou logickou ochranu takto vzniklých digitálních dokumentů. Není tím myšlena ochrana dat ve smyslu zálohování (ochrana bitstreamu), ale zajištění použitelnosti současných digitálních dokumentů v budoucnu. K tomu, aby logická ochrana mohla být prováděna, je potřeba zapojit do životního cyklu digitálních objektů nové nástroje a vytvářet nové typy metadat. I u vyspělých systémů na ukládání dat logická dlouhodobá ochrana stojí a padá s metadaty, která jsou v systému uložena, nebo která ten konkrétní systém vytváří, doplňuje, spravuje. Jde nejen o metadata popisná, ale především o strukturální, administrativní, ochranná a technická. Právě o těchto metadatech je předkládaná práce.

V ČR se digitalizace provádí již od poloviny 90. let minulého století. Ovšem nikdy nevznikla specifikace metadat, která by napomáhala logické dlouhodobé ochraně takto vzniklých digitálních objektů, a to i přesto, že tato problematika je od roku 2005 velmi aktuální a je řešena na mezinárodních fórech a od roku 2005 vznikaly také relevantní metadatové standardy (např. PREMIS). Určitým pokusem o zavedení nových typů metadat do procesu digitalizace bylo v roce 2008 doplnění specifikace metadat pro program VISK7 doplněna o omezený počet elementů schémat PREMIS a MIX.

Téma disertační práce jsem si vybral právě s ohledem na výše uvedené skutečnosti. Dalo se předpokládat, že bude nutné přijít s novým návrhem na tvorbu metadat v procesech digitalizace v NK ČR, který by reflektoval potřebu vytváření administrativních, ochranných a technických metadat tak, aby bylo možné je použít pro dlouhodobou ochranu digitálních dat ve speciálních systémech pro uložení dat. Z tohoto důvodu bylo nutné popsat vývoj a stav využití metadat v obou hlavních projektech NK ČR, které se zabývají digitalizací (*Kramerius* a *Manuscriptorium*). Popis vývoje je východiskem návrhu nových profilů metadat. Příležitostí, jak se k této problematice postavit čelem a v plné míře, byl projekt *Národní digitální knihovna*. Vznikl návrh nové podoby metadat pro masovou digitalizaci v projektu NDK, který reflektuje poslední vývoj v oblasti a obsahuje metadata nejen popisná, ale i ochranná, technická a administrativní. Tento návrh je součástí předkládané disertační práce.

Proces digitalizace vždy začíná otázkami. Jaké standardy dat a metadat vybrat, aby splňovaly naše potřeby a možnosti? Kam a jakým způsobem budeme ukládat naše data? Jak bude vypadat naše workflow? Jaké HW vybavení použijeme? Jaké můžeme očekávat problémy během digitalizace a řešil je již někdo? Budeme digitalizovat sami nebo budeme proces zadávat externí firmě? Tato práce se z valné většiny věnuje prvním dvěma otázkám, tedy standardům metadat pro digitalizaci a možnostem uložení, které s metadaty přicházejícími z digitalizace úzce souvisejí.

2.1 Vymezení používaných termínů

aplikační metadatový profil – deklarace metadatových schémat a jejich elementů, které organizace, aplikace, komunita používá pro tvorbu svých metadatových záznamů; sada metadatových elementů, strategií a návodů definovaných pro konkrétní implementaci v konkrétním systému; elementy mohou být z jednoho nebo více sad elementů (*metadata set*) zkombinované v jedno výsledné schéma nebo implementaci více schémat v kontejnerovém standardu (např. METS)

archivní kopie – digitální objekt určený k archivaci; jde o maximálně kvalitní digitální verzi předlohy, nebo digital-born dokument; není určen pro uživatele a z tohoto důvodu se z ní často vytváří tzv. uživatelská kopie; archivní kopii se věnuje v LTP systému veškerá pozornost a prochází všemi kroky na vstupu do repozitáře (validace, identifikace formátů, charakterizace), následně pak plánováním ochrany a ochrannými akcemi

atribut elementu metadat – bližší určení (charakteristika) elementu nebo entity; pro oblast XML sestává z páru jméno – hodnota, odděleného rovnítkem; může se vyskytovat pouze uvnitř tagu XML elementu [KAŠPAROVÁ a PSOHLAVEC, 2008, s. 6]

autenticita digitálního objektu – digitální objekt je považován za autentický, pokud je možno prokázat, že odpovídá předloze a že se od svého vzniku nezměnil, nebo se změnil s vědomím jeho správce a tato změna je zdokumentována včetně jejího zdůvodnění a výsledku; autenticitu pomáhají udržovat a dokumentovat metadata o událostech a vzniku digitálního objektu

data – „*Reprezentace informací vhodně formalizovaná pro komunikaci, interpretaci a zpracování lidmi a automaty. Data mohou být reprezentována libovolnými řetězci znaků (čísel, příkazů, vět) uloženými na informačním nosiči*“ [JONÁK, 2003]; pro účely této práce budeme vycházet z definice dle modelu životního cyklu organizace *Digital Curation Centre*, který uvádí, že jde o jakoukoliv informaci v binární digitální formě [DIGITAL CURATION CENTRE, 2010]

digital-born dokumenty – digitální objekty (dokumenty, materiály), které vznikly v digitální podobě za použití počítače a nadále v této podobě existují; digital-born dokument může také být použit na výrobu analogové kopie, např. vytištěním

digitalizace – převod analogových dokumentů do digitální podoby

digitalizovaný dokument – výsledek procesu digitalizace

digital curation – zabývá se celým životním cyklem digitálního dokumentu a směřuje k jeho dlouhodobému uchování a použitelnosti; jednou z podčástí je i logická dlouhodobá ochrana digitálních dat (*digital preservation*)

digitální knihovna – termín nahrazuje starší termíny elektronická nebo virtuální knihovna a může mít tři různé významy: 1) knihovna, jako instituce, která část svých sbírek zpřístupňuje nebo ukládá v digitální, strojem čitelné podobě; digitální data mohou být uložena přímo v budově knihovny nebo mimo ni, jsou však vždy součástí sbírek; 2) digitální knihovnou může být také někdy nazývána aplikace zpřístupnění, která jako nadstavba repozitáře umožní uživateli vyhledávání, dodání a prohlížení digitálních objektů[3]; 3) digitální knihovnou se také může rozumět celek skládající se z digitálního repozitáře, aplikace zpřístupnění a digitálních dokumentů, tedy informační systém pro vkládání, uchovávání, organizaci, správu, vyhledávání a zpřístupňování digitálních materiálů uživatelům

digitální objekt

- jednoduchý digitální objekt je diskrétní (nespojitá) digitální jednotka; odpovídá jednomu počítačovému souboru, může jít o obrazový (např. JPEG), textový (TXT) nebo jiný soubor, případně doprovázený metadaty
- komplexní digitální objekt – je diskrétní (nespojitý) objekt skládající se z jiných digitálních objektů jednoduchých; v kontextu digitalizace a této práce označuje nejčastěji množinu naskenovaných dat (nejčastěji digitální obrazy) a k nim náležejících metadat, která tvoří dohromady digitální intelektuální entitu (digitální kopii fyzického dokumentu)

digitální repozitář – celek složený z HW pro uložení dat, ze SW aplikace ke správě dat a správců (osoby, SW); řešení pro uložení digitálních objektů; někdy se v češtině používá rovnocenný termín „digitální úložiště“, tento termín byl používán v souvislosti s konkrétním repozitářem NK ČR (Centrální datové úložiště) a je v tomto kontextu zachován

dokument – informační pramen tvořený nosičem informací a množinou informací na něm fixovaných a sloužící k přenosu dat v čase a prostoru; dokumenty lze dělit dle různých kritérií, např. podle způsobu záznamu dat (písemné, obrazové, zvukové, elektronické/digitální), podle kontinuity (periodické a neperiodické) apod. [KAŠPAROVÁ a PSOHLAVEC, 2008, s. 7]

DTD (Document Type Definition) – ve formátech SGML a XML formální popis součástí určitého dokumentu nebo třídy dokumentů; DTD je strojově čitelný soubor popisující pomocí formální syntaxe, které elementy a entity se mohou v dokumentu vyskytovat a kolikrát, a jaký je povolený obsah a atributy těchto elementů (lze přirovnat ke slovníku a mluvnickým pravidlům); příkladem DTD je definice jazyka HTML pro vytváření dokumentů určených pro službu WWW či různá schémata pro popis digitálních zdrojů (TEI, EAD) [CELBOVÁ, 2003-]; DTD soubor umožňuje nejen vytváření metadatových záznamů podle určitých pravidel, ale i validaci těchto metadatových záznamů oproti konkrétnímu DTD záznamu; validací lze zjistit správné použití elementů a atributů v metadatovém popisu

[3] Takto je definovaná i v *Manifestu digitálních knihoven* z projektu DELOS [CANDELA et al., [2006]].

digitální (elektronický) dokument – dokumenty kódované pomocí binárních znaků (0 a 1); elektronické dokumenty nejsou vázány na žádný nosič, mohou být uložené na fyzickém nosiči (optický disk, pevný disk) nebo existovat pouze online; výrazy „digitální dokument“ a „elektronický dokument“ jsou totožné

element metadat – pro oblast XML a metadat se jedná o základní složku obsahu metadatového záznamu; skládá se ze startovního tagu a koncového tagu a obsahu elementu – jeho hodnoty; rozlišujeme element rodičovský (nadřazený) a dceřiný (podřízený), který je do rodičovského vnořen

fyzický dokument – originální předloha složená z jedné nebo více částí, které tvoří dohromady logický celek určený např. k digitalizaci

historické fondy – knihovní rukopisné nebo tištěné fondy vzniklé do 19. století, tj. do roku 1800

historický dokument – pro potřeby této práce označujeme tímto pojmem všechny fyzické písemné dokumenty uložené v knihovnách vzniklé do roku 1800; rozlišujeme mezi rukopisem, inkunábulí (dokument vytištěný v období od vynálezu knihtisku do roku 1500) a starým tiskem (dokumenty vytištěný v letech 1501 – 1800)

integrita jednoduchého digitálního objektu – zachování původní podoby bitstreamu tvořícího digitální objekt; jakákoliv změna bitů digitálního objektu může způsobit jeho poškození a znemožnit jeho použití, zpřístupnění; změna bitů může být ovšem chtěná, a to např. při migraci do jiného formátu, migrací ovšem vzniká jiný digitální objekt; jedná se o fyzickou integritu digitálního objektu

integrita komplexního digitálního objektu – zachování původní skladby komplexního objektu tvořené dalšími jednoduchými digitálními objekty; pokud je jeden objekt změněn a např. má novou podobu díky migraci do nového formátu, integrita digitálního dokumentu jako celku není narušena; narušena je pokud některý z původních objektů chybí, je poškozen nebo není např. jasný původ objektu; integrita z pohledu dlouhodobé ochrany je širší pojem a zahrnuje mj. důvěryhodnost digitálního objektu, kontext, údaje o původu, autenticitě, celistvosti (*fixity*), formu obsahu a reference na digitální objekt; lze tedy říci, že jde o intelektuální integritu komplexního digitálního objektu

intelektuální entita – dle datového modelu PREMIS celek, který lze považovat za logický, lze popsat, samostatně prezentovat; může to být kniha, číslo časopisu apod.; entita může mít fyzickou podobu (fyzický dokument) i digitální podobu (komplexní nebo i jednoduchý digitální objekt) a může se skládat z jiných entit

interoperabilita (metadat) – schopnost různých systémů s různými HW a SW platformami, datovými strukturami a rozhraními vyměňovat data a metadata s minimální ztrátou obsahu a funkcionality [NISO, 2004, s. 2]; v kontextu metadat se interoperabilitou rozumí vlastnost dvou

nebo více systémů, která umožňuje bezproblémové využívání, výměnu, přejímání, zpřístupnění různých metadat (a tedy i dokumentů) mezi těmito systémy bez větších problémů

komplexní digitální dokument – výraz používaný v prostředí digitální knihovny Manuscriptorium; označuje kompletní množinu dat a metadat, která tvoří intelektuální entitu (rukopis, monografii, apod.) anebo ji doplňují; doplněním je myšleno např. to, že k digitální kopii určité intelektuální entity (rukopis) jsou přidány intelektuální entity další s ní související (audio záznam, článek, plný text aj.) – více viz [AIP Beroun, 2005]; v pojetí projektu *Manuscriptorium* může nastat situace, kdy komplexní digitální dokument existuje, aniž by fyzická předloha byla zdigitalizována; nejčastěji v případě, že jde o záznam katalogu historických fondů v Manuscriptoriu; záznam tohoto katalogu obsahuje pouze popisná metadata a to ve stejném formátu jako má komplexní dokument

Kramerius

- a) *projekt Kramerius*; původně označení národního programu ochranného mikrofilmování, který začal v roce 1992; hlavním cílem byla obnova mikrografických pracovišť v českých knihovnách, optimalizace ochranného mikrofilmování dle ISO norem a vybudování pracoviště pro skenování mikrofilmů [POLIŠENSKÝ, 2008, s. 51]; projekt byl od počátku zaměřen na ochranu dokumentů ohrožených kyselostí papíru, později byl doplněn o možnost vytváření digitálních kopií digitalizací mikrofilmů (projekt *Digitalizace mikromédií* 1997-1999) a také pomocí *hybridní metody* (1999); dnes je primárním výstupem digitální kopie dokumentu (buď z papírové předlohy, nebo z mikrofilmu); od roku 2000 je *Kramerius* součástí programu *Veřejné informační služby knihoven* (VISK) jako VISK7
- b) *digitální knihovna* NK ČR, obsahuje monografie a periodika, tedy výstupy z digitalizace NK ČR samotné a všech institucí účastnících se programu VISK7
- c) *open source (GNU GPL) CMS* (*Content Management System*); programová aplikace sloužící ke zpřístupnění digitálních dat na Internetu; vznikla na popud knihoven, které v rámci VISK7 digitalizovaly a požadovaly možnost zpřístupnění na lokální síti nebo na Internetu, namísto CD-ROM; pomocí aplikace lze provádět import, export dat, mazání, úpravy digitálních dokumentů ve vrstvě zpřístupnění, replikace dat do dalších systémů Kramerius, uživatelské vyhledávání atd.; původní aplikaci vyvinula firma Qbizm v rámci VaV *Vytvoření virtuálního badatelského prostředí pro zpřístupnění a ochranu digitálních dokumentů* (2004-2010); od roku 2010 existuje kompletně nová verze Kramerius4 (někdy také K4), kterou vyvíjí firma INCAD; více o vývoji a technickém řešení systému do verze 3 viz např. [LJUBKA, 2008] nebo [LHOTÁK, 2007]; v textu práce je systém Kramerius často uváděn jako „aplikace zpřístupnění“

Manuscriptorium – název projektu, systému i vlastní digitální knihovny pro zpřístupnění historických dokumentů vzniklých/vydaných do roku 1800; systém Manuscriptorium vznikl v roce 2003 jako otevřený katalog historických fondů a posléze i jako digitální knihovna s digitálními dokumenty (vývoj AiP Beroun) pod názvem Memoria; primární zaměření digitální knihovny je na rukopisy, inkunábule, staré tisky, historické mapy; digitální knihovna zpřístupňuje údaje o historických fondech jednotlivých spolupracujících institucí, jejich digitální obrazy i plné texty (hlavně výstupy projektů *Memoriae Mundi Series Bohemica* a VISK6); obsah dodává nejen NK ČR, ale také další paměťové instituce – knihovny, muzea, archivy z ČR i ze zahraničí

mapování metadat – sémantické mapování elementů a atributů mezi dvěma nebo více metadatovými schématy navzájem; tj. elementy a atributy jednoho schématu vůči elementům schématu jiného

markup (značkovací jazyk) – poskytuje slovník a syntax ke značkování prostého textu; značky (tagy) textu dodávají smysl a jsou vodítkem pro strojové zpracování různého druhu

metadata – strukturovaná data, která nesou informace o primárním dokumentu a slouží k jeho popisu; pojem metadat je používán především v souvislosti s elektronickými zdroji

metadatová specifikace – někdy také datový slovník metadat; rozšířený popis metadatového schématu

metadatové schéma – někdy také metadatový formát; z technického pohledu jde o strojově zpracovatelné specifikace definující seznam elementů a jejich atributů a jejich struktury ve formálním jazyce (ve formě DTD nebo nověji XSD souborů); z pohledu obecného jde o specifikaci možné struktury záznamu, syntaxe a sémantiky elementů včetně pravidel pro jejich plnění; obecně je metadatové schéma „*sada metadatových elementů určená ke konkrétním účelům, jako např. k popisu konkrétního typu informačního zdroje*“ [NISO, 2004, s. 2]

metadatový standard – metadatové schéma, které se stane normou (NISO, ISO aj.) a je široce využíváno; může jít také o tzv. de-facto standard – není opravdovou normou, ale je za ni považován

metadatový záznam – konkrétní popis určitého digitálního nebo analogového objektu, který je strukturován a vytvořen z elementů podle konkrétního metadatového schématu a zapsán nejčastěji v XML nebo jinak; jde tedy o XML nebo textový soubor; metadatový záznam je považován za základní jednotku správy a výměny mezi systémy

migrace

- a) migrace formátová – změna formátu výchozího digitálního objektu na jiný formát cílový, např. doc > pdf
- b) migrace fyzická – na nový HW nebo fyzický nosič ve smyslu kopírování z jednoho HW nebo nosiče na jiný, většinou novější při výměně technologií
- v kontextu této disertační práce termín „migrace“ bez dalšího vysvětlení označuje vždy migraci formátovou

novodobé fondy – knihovní fondy vzniklé od 19. století (1800) do současnosti

novodobý dokument – pro potřeby této práce označujeme tímto pojmem veškeré fyzické písemné dokumenty uložené v knihovnách vzniklé po roce 1800 do současnosti (primárně knihy, periodika)

obraz (obrazový dokument) – v kontextu této práce se rozumí statický obrazový dokument např. ve formátu JPG, JPEG 2000, TIFF aj.; záměrně není využit termín obrázek

ochranná akce (aktivita) – opatření dlouhodobé ochrany vyplývající z rizik a plánů ochrany vedoucí ke snížení míry rizika nebo jeho odstranění (migrace, emulace aj.)

ochranné reformátování – digitalizace tradičních dokumentů za účelem jejich ochrany a zpřístupnění

plán dlouhodobé ochrany digitálních dat – definuje sled ochranných akcí, které provádí instituce zabývající se dlouhodobou ochranou digitálních dat v reakci na identifikovaná rizika u svých digitálních objektů; plán bere v potaz strategii ochrany, omezení vyplývající z legislativy, nároky uživatelů a cíle ochrany; popisuje kontext ochrany digitálních dat, použité strategie i důvody provádění ochranných akcí

plánování ochrany – schopnost odhadnout výskyt a případný dopad rizik a specifikovat podle těchto skutečností plán ochrany, který definuje konkrétní směr ochranných aktivit a podmínky jejich provedení; jde vlastně o předcházení zastarávání formátů, problémů s digitálními objekty apod.

registr formátů – online služba poskytující seznam běžně užívaných formátů, který obsahuje pro každý formát dat identifikátor, popis vlastností, rizik s formátem spojených, odkazy na případnou dokumentaci apod.; registry hrají důležitou úlohu pro funkčnost LTP systémů a pro automatické obohacování digitálních objektů o technická a administrativní metadata; nejznámějšími zástupci jsou PRONOM[4] a UDFR[5].

SGML – standardní obecný značkovací jazyk; norma ISO 8879 z roku 1986 pro strukturování elektronických textů; SGML byl prvním „meta" značkovacím jazykem; vyvinut jako platforma pro vývoj dalších značkovacích jazyků a jejich DTD; poskytuje pravidla pro pojmenování elementů a také syntax pro vyjádření logických vazeb mezi jednotlivými částmi dokumentu; pomocí tzv. definic typu dokumentu (DTD) lze definovat vlastní strukturní značkovací jazyk, např. jazyk HTML je definován pomocí SGML [SKLENÁK, 2003], tedy HTML je DTD derivace ze SGML; SGML neomezuje nijak počet elementů ani jejich atributů, ani jejich pojmenování

tradiční dokument – analogový (fyzický) dokument, který se tradičně objevuje v knihovnách; obsahuje informace ve spojité formě a nelze jej proto kopírovat bez ztráty kvality/informací; analogové dokumenty jsou nutně spojeny s určitým typem nosiče, např. papírové knihy, periodika, plakáty, mapy, grafika nebo nepapírové dokumenty jako např. audio kazety, video kazety; může jíž o strojem čitelný dokument, ovšem nemusí být potřeba počítač k jeho dekódování

uživatelská kopie – verze digitálního objektu, která je určena uživatelům; většinou se jedná o derivát archivní (master) kopie, který je upravený tak, aby byl vhodnější pro uživatele z hlediska použití (menší velikost souboru, odlišný formát digitálního objektu); může vznikat již

[4] http://www.nationalarchives.gov.uk/pronom/
[5] http://www.udfr.org/

při digitalizaci, nebo na vyžádání (*on-the-fly*) přímo v repozitáři v okamžiku, kdy si dokument uživatel žádá

VISK – *Veřejné informační služby knihoven*; program financování různých oblastí knihovnických a informačních činností včetně podpory nekomerčních projektů z oblasti knihoven; hlavním cílem programu připravovaného Ministerstvem kultury ČR ve spolupráci s NK ČR a Svazem knihovních a informačních pracovníků (SKIP), je „*inovace veřejných knihovnických a informačních služeb na bázi informačních technologií (ICT)*" [ČESKO. MINISTERSTVO KULTURY, 2005]; program je výsledkem *Akčního plánu realizace státní informační politiky*, jak jej přijala vláda ČR v roce 1999 a byl vyhlášen 10. dubna 2000 usnesením vlády č. 351 o *Koncepci státní informační politiky ve vzdělávání*[6]. Program VISK má 9 podprogramů, z nichž osmý je rozdělen na dva. Digitalizace se bezprostředně dotýkají:

- VISK4 – Digitální knihovna a archiv pro informační služby knihoven; zaměřen na podporu aktivit VISK6 a VISK7, vývoj aplikací pro zpřístupnění digitálních dat (aplikace digitálních knihoven Kramerius a Manuscriptorium); třetí částí je oficiálně od roku 2009 podpora Centrálního datového úložiště NK ČR, na kterém jsou uložena data pro obě výše zmíněné digitální knihovny, včetně dat ostatních knihoven zapojených do VISK7
- VISK5 – Národní program retrospektivní konverze katalogů knihoven v ČR – RETROKON; cílem je zpřístupnění knihovních katalogů prostřednictvím Internetu pomocí přepisu lístků do databáze katalogu, nebo alespoň jejich převodem do digitální podoby, tj. digitalizace – viz obrazový katalog KATIF[7]
- VISK6 – Národní program digitálního zpřístupnění vzácných dokumentů *Memoriae Mundi Series Bohemica*; základním cílem programu je zajistit metodou digitalizace ochranu a široké zpřístupnění vzácných dokumentů knihoven a dalších sbírek tvořících důležitou součást národního kulturního dědictví [KNOLL, 2011b]; výstupem projektu jsou digitalizované dokumenty (rukopisy, staré tisky a mapy) dostupné v digitální knihovně Manuscriptorium, dále specifikace metadatových formátů, metody digitalizace historických dokumentů aj.
- VISK7 – Národní program mikrofilmování a digitálního zpřístupňování dokumentů ohrožených degradací kyselého papíru – *Kramerius*; cílem programu je záchrana a zpřístupnění bohemikálních dokumentů tištěných na kyselém papíře, jejichž existence je ohrožena rozpadem (křehnutím) papírového nosiče; původním primárním výstupem byly dokumenty na mikrofilmu, později přibyla možnost digitalizace těchto mikrofilmů a v současné době je primárním výstupem digitální kopie; dokumenty vzniklé v rámci VISK7 archivuje NK ČR a zároveň je zpřístupňuje ve své digitální knihovně Kramerius.

XML záznam (dokument) – uspořádaný a řádně označený stromovitý záznam, který obsahuje elementy (tagy); jejich obsahem může být cokoliv, co lze reprezentovat pomocí kódu Unicode; XML záznam může obsahovat i řetězce bitů, reprezentující např. digitální obraz nebo jiný digitální objekt

[6] Více viz web projektů VISK http://visk.nkp.cz/.
[7] http://katif.nkp.cz

3. Digitalizace, metadata a dlouhodobá ochrana digitálních dat

Digitalizace je laickým pohledem skenování knih. Odborněji je možné říci, že jde o převod analogových dokumentů do digitální podoby. Mezi hlavní důvody digitalizace patří ochrana fyzických předloh a také zpřístupnění, které začíná v posledních letech převládat. Laická veřejnost většinou nechápe komplexnost procesu digitalizace, který nezahrnuje pouze převod do digitální podoby, ale také vytváření metadat, které zabírá více času než samotné skenování. Návazné procesy jsou uložení, správa dat a zpřístupnění. Digitalizace tak velmi souvisí s metadaty i s dlouhodobou ochranou digitálních dat, která je dnes považována za součást digitalizace. Jinými slovy, jak nastavíme digitalizaci, takové máme možnosti dlouhodobé ochrany dat a jejich zpřístupnění v budoucnu.

Digitalizace v paměťových institucích se začala ve větší míře objevovat v 90. letech 20. století (viz např. program *Paměť světa* podporovaný organizací UNESCO). Opravdový impulz pro digitalizaci znamenal příchod a rozšíření Internetu (srovnej např. s [BÜLOW a AHMON, 2011]). Internet se stal hlavním zdrojem různých typů informací pro miliony lidí. Změnil nejen náš život, ale i digitální knihovny a knihovny samotné, na které klade nové nároky v oblasti zpřístupnění a dostupnosti dokumentů. Lidé chtějí získávat informace online a to co nejsnáze. Digitalizace nabídla novou možnost zpřístupnění. Namísto jednoho uživatele studujícího jeden konkrétní fyzický dokument na konkrétním místě (budova knihovny), může digitální podobu tohoto dokumentu studovat takříkajíc neomezený počet uživatelů a to odkudkoliv a kdykoliv. Ve stejné době, kdy se rozvíjel a rozšiřoval Internet, došlo také k rozvoji technologií, které jsou pro digitalizaci klíčové a dovolují digitalizovat i menším knihovnám nebo archivům (skenovací technologie, ukládání dat, sdílení po síti, dostupnost PC apod.). Digitalizace navíc umožňuje sdílení, agregování a kombinování dokumentů, které by v analogovém světě nebylo myslitelné (např. tematické portály, smíšené digitální knihovny více institucí, agregátory jako *Europeana* apod.). Zpřístupnit lze i dokumenty, které byly v analogové podobě ve stavu, kdy uživatelům nemohly býti poskytnuty. A naopak, uživatelé díky vyhledávání často nalézají dokumenty, které v papírové podobě byly minimálně využívané. Ukázalo se, že zdigitalizované obrazy poskytují možnosti zpřístupnění, o kterých se nikomu před pár lety ani nesnilo.
Možnosti digitalizace a technologie samotné mění zaběhlé postupy a cíle paměťových institucí. To lze nejlépe vidět např. na tom, jak tyto instituce mění a upravují své cíle i koncepce. Ve většině z nich se objevuje alespoň zmínka o digitalizaci, u největších z nich je digitalizace mnohdy jedním z hlavních bodů/cílů – viz strategie Britské knihovny na léta 2008-2011, která má za cíl zpřístupnit uživatelům veškeré znalosti pomocí sedmi priorit. Pět z oněch priorit má vylepšit digitální infrastrukturu Britské knihovny a využívání digitálních dokumentů, které knihovna má [BRITISH LIBRARY, 2008].

3.1 Důvody a přínos digitalizace

Naši současnou společnost lze nazvat informační nebo znalostní. Jde o společnost, která využívá elektronické dokumenty a zdroje a vyžaduje mj. i elektronickou podobu analogových zdrojů. Tato potřeba je dnes hlavním motorem digitalizace a do jisté míry vytlačila původní cíl digitalizace, kterým byla ochrana a uchování obsahů analogových dokumentů. Digitalizace je jednou z několika možností tzv. reformátování, tj. přenosu obsahu na jiný nosič, než je ten původní. První z metod

reformátování bylo mikrofilmování, které se vyvíjelo již od konce 19. století a svůj vrchol zažilo ve druhé půli století dvacátého. Digitalizace začala mikrofilmování doplňovat na konci 20. století (*hybridní metoda*[8]) a v novém tisíciletí jej začala nahrazovat zcela (zrušení mikrografických aktivit v národních knihovnách v zemích jako jsou např. Nizozemí, Velká Británie apod.).

Digitalizace je ukázkovým příkladem vhodného řešení věčného rozporu knihovnictví, které má za úkol dokumenty v maximální míře zpřístupňovat a zároveň je chránit pro další generace.[9] Tento rozpor byl markantní v případě preventivní ochrany fyzických dokumentů, která se se zpřístupňováním příliš neslučovala. A právě digitalizace je nástrojem, který velmi dobře podporuje obě uvedené aktivity a tento rozpor do jisté míry smazal. Jako jedna z možností reformátování byla digitalizace nutností pro záchranu ohrožených dokumentů, zároveň poskytuje možnost vystavení uživatelských kopií širokému spektru čtenářů, aniž by se ničila původní předloha. Z tohoto popudu se začalo s digitalizací v NK ČR, kdy první digitalizovaný rukopis vytvořený v roce 1993 pro program UNESCO *Memory of the World*, byl myšlen jako náhrada za originál pro prezenční výpůjčky ve studovnách rukopisů a starých tisků.

Ovšem předpoklad, že digitální kopie ochrání fyzickou předlohu před půjčováním a tedy opotřebením, není zdaleka tak nezpochybnitelný. Existují názory a důkazy, že tomu je mnohdy přesně naopak. Tedy, že zpřístupnění online verzí analogových dokumentů přivádí zájem více uživatelů k jejich fyzickým předlohám. Např. Národní archiv Velké Británie měl ke konci roku 2009 online přístupných asi 80 milionů dokumentů. Počet stažených/prohlédnutých obrázků za jeden měsíc byl 12 milionů. I přesto se ukázalo, že počet dokumentů půjčených fyzicky ve studovnách neklesá, ale zůstává stejný [BÜLOW a AHMON, 2011, s. 8].

Dnes není problémem proces digitalizace, ten lze považovat za daný a do jisté míry jasný. Aktuální problém paměťových institucí celého světa je otázka, jak zachovat digitalizované a digital-born dokumenty použitelné i pro budoucnost. Digitalizované dokumenty a s nimi i tzv. digital-born dokumenty se samy staly objektem ochrany. Tj. problém, který digitalizace řešila, se objevuje znovu, ovšem můžeme říci, že v komplikovanější podobě. Objektem ochrany jsou totiž digitální data.

Ne vždy je také digitalizace všeobjímajícím řešením pro zpřístupnění. Knihy a dokumenty v jakékoliv historické době vznikaly převážně k zachycení intelektuálního obsahu (až na výjimky, jako jsou určité historické rukopisy). Badatele tedy nejvíce zajímá obsah, tj. psaná a reprodukovatelná část stránek. V dnešní době je ale mnoho z těchto historických svazků zajímavých i jako artefakt. Existuje významná skupina badatelů a odborníků (např. paleografů, kodikologů apod.), kteří z různých důvodů budou vždy pro svá bádání požadovat originál dokumentu, například pro podrobné zkoumání psacích látek, použitých materiálů, kvůli skladbě knihy atp.

V případě historických dokumentů, kdy často jde o jedinečné jednotliviny s nádhernými vazbami a uměleckým provedením, digitální kopie takový originální artefakt může velmi přiblížit, plně nahradit jej však nemůže nikdy. Zcela reprodukován je obsah dokumentu, ne ovšem jeho původní

[8] Metoda vytváření digitální kopie a mikrofilmu v jediném okamžiku.

[9] V našem prostředí jde o povinnost vycházející z knihovnického zákona (zákon č. 257/2001 Sb.), který mluví o zajištění ochrany pro písemné dědictví uložené v knihovních sbírkách. V § 18 říká, že provozovatel knihovny je povinen zajistit vhodné uložení knihovního fondu, ochránit jej před poškození a nepříznivými vlivy prostředí, zajistit restaurování dokumentů v případě potřeby nebo jejich převedení na jiný druh nosič, je-li to potřeba k trvalému uchování dokumentů.

trojrozměrná podoba. Slovy Zdeňka Uhlíře: „*Zachovány zůstávají rysy paleografické …, zanikají rysy kodikologické.*“ [UHLÍŘ, 1999a, s. 118] Z původního trojrozměrného dokumentu se digitalizací stal dokument dvojrozměrný. Digitální kopie mohou v určitých ohledech badateli pomoci, pokud nabízejí funkčnosti, které při studiu fyzického dokumentu nejsou možné, jako např. různé úhly osvětlení stránky, úpravy kontrastu, negativní zobrazení apod. Nevýhodou zůstává, že při digitalizaci, ať sebekvalitnější, se určité procento informace z původní předlohy ztrácí. U novodobých dokumentů je koncept digitalizace o tuto starost jednodušší. Novodobé dokumenty zpravidla vznikaly a vznikají sériovou výrobou, není tedy nutné tak podrobně dokumentovat jejich fyzickou podobu.

3.1.1 Digitalizace jako ochrana fyzických knihovních fondů

Chápání digitalizace jako odpovídající formy ochrany fyzických fondů ovšem nebylo tak přímočaré a rychlé. Správci sbírek a experti na ochranu knihovních i jiných fondů od počátku měli, a dodnes mají, velké obavy spojené s tím, že digitalizací se sice ochrání předloha proti nadměrnému užívání a je ochráněn intelektuální obsah v případě ztráty předlohy, ale samotný digitální objekt je daleko méně stabilní než papír, pergamen a jiné další psací látky tradičně používané. H. Weber a M. Dorr se ve své zprávě k této problematice z roku 1997 ptají v úvodu: „*Jak se má zacházet z hlediska ochrany s médiem, které je notoricky známé svou nestálostí a pro které je deset let dlouhodobý horizont? Jaký má smysl spoléhat se na takovou technologii, když řešíme záchranu papírových materiálů, které pomalu degradují 100 a více let?*“ [DORR a WEBER, 1997]

Původní obava byla spojena s životností nosiče digitálního dokumentu, který byl obecně považován za velmi nestálý (optické disky apod.) a tudíž za dlouhodobě nevhodný. Postupem času se ukázalo, že daleko větším problémem, který může digitalizaci diskvalifikovat z okruhu uznávaných metod tzv. ochranného reformátování (ochrana fyzických předloh), je závislost nosiče i samotného digitálního dokumentu na konkrétní technologii (zastarávání SW a HW). Zvláště markantní byl tento problém u zastarávání SW, se kterým byly digitální objekty často spojeny z výroby nebo v oblasti zpřístupnění. Za řešení ochrany samotných digitálních objektů byla navrhována vedle emulace také migrace formátů. Ovšem ještě v roce 1999 Abby Smith ve svém, v knihovnické komunitě velmi známém, článku „*Why Digitize?*“ píše, že neexistuje žádné ověřené řešení tohoto problému [SMITH, 1999, s. 6]. Stejně závažnou překážkou tomu, aby bylo možno prohlásit digitalizaci za uznanou metodu ochranného reformátování, byla nemožnost prokázat u digitálních objektů jejich integritu a autenticitu, tj. jejich případné úpravy, změny apod.[10] Ze všech těchto důvodů bylo za odpovídající způsob reformátování považováno mikrofilmování. U mikrofilmů se běžně uvádí životnost v běžných podmínkách okolo 500 let, což bylo pro digitalizaci a digitální objekty obtížně představitelné.[11]

[10] Autenticita dokumentu z hlediska uživatele a správce zahrnuje identitu (dokument je tím, za co se vydává) a integritu (kompletnost dokumentu). Autenticita je ohrožena, protože paměťové instituce neupravují své procesy na digitální objekty. Nikdo neřekne archivářům, že nesmí digitální objekty měnit, a když je změní, že musí změny zaznamenat i s jejich důvody a výsledky. U digitálních dokumentů je změna navíc neviditelná, často není možné zjistit, že změna byla vůbec provedena.

[11] Mikrofilmování zažilo svůj boom jako forma ochrany a náhrady fyzické předlohy po roce 1954, kdy byla schválena Haagská konvence (Konvence o ochraně kulturního dědictví v ozbrojeném konfliktu). V konvenci se nabádalo k přípravám k ochraně v době míru, které mohou pomoci předejít ztrátám v době války. Jako forma přípravy byl přijat mikrofilm v roli náhrady eventuálně poškozené předlohy. Již v té době bylo mikrofilmování zvládnutý proces se zavedenými standardy a materiály.

Pocit nejistoty spojené s digitalizací zesílil v okamžiku, kdy v mnoha knihovnách převládlo nadšení nad digitalizací do té míry, že finanční prostředky určené na ochranu fyzických fondů začaly být směřovány právě na digitalizaci. Výše uvedená zpráva [DORR a WEBER, 1997] vznikla z popudu *Deutche Forschungsgemeinschaft* (DFG), která viděla možnost řešení problému v kombinaci digitalizace a mikrofilmování, tedy v hybridní metodě. Zpráva zevrubně popisuje zásady a kvalitativní požadavky jak na digitalizaci, tak na mikrofilmování. Hybridní metoda měla opravdu své opodstatnění a byla na přelomu tisíciletí široce využívána, také v NK ČR, která dodnes vlastní hybridní kameru – více viz [POLIŠENSKÝ, 2000].

Postupem času, s vývojem systémů na uložení a později na dlouhodobou ochranu digitálních dat (tzv. LTP systémy), se situace měnila a digitalizace začala být považována za právoplatný způsob ochrany fyzických dokumentů. Od 90. let 20. století vznikaly návody na správný postup digitalizace, velmi často spolu se specifikacemi na mikrofilmování v jednom dokumentu. I přesto ještě v roce 2002 v reprezentativním průzkumu provedeném Britskou knihovnou zúčastnění odborníci odpovídají, že digitální podobu dokumentů nepovažují za vhodnou pro dlouhodobou ochranu v porovnání s papírem. Digitální podoba dokumentů byla považována za vhodnou pro zpřístupnění, ale jako archivní médium byly digitální objekty považovány za: „*nestabilní, nedospělé, nevyzkoušené ve větším množství a nespolehlivé z dlouhodobého hlediska.*“ [SHENTON, [2004]] První publikace, která na základě určitého výzkumu a hledisek uváděla, že digitalizace je odpovídající formou reformátování, stejně jako např. mikrofilmování, byla americká zpráva od *The Association of Research Libraries* (ARL), která v roce 2004 vydala publikaci s názvem *Digitization as a preservation reformatting method* [ARTHUR, 2004]. Zpráva obsahovala souhrn aktivit v oblasti dlouhodobé ochrany – vývoj na poli metadat, SW pro repozitáře, nástrojů, seznam best practices pro digitalizaci, seznam institucí aktivních na tomto poli, seznam informačních zdrojů a další údaje, které upevňovaly postavení dlouhodobé ochrany digitálních dat jako svébytné tématiky, řešení na všech polích a frontách národních i mezinárodních. Zpráva obsahovala také detailní porovnání výhod a nevýhod mikrofilmů a digitálních kopií [ARTHUR, 2004, s. 8]. Tato zpráva působila jako „urychlovač“ procesu přijetí digitalizace jako formy ochrany, i když takovéto chápání digitalizace nepřišlo hned. V následujících letech se ale dostalo do bodu, kdy digitální data začala být považována za odpovídající náhradu fyzické předlohy i pro budoucnost [WALKER, 2007, s. 514].

Digitální objekty lze dnes ukládat do tzv. systémů na dlouhodobou ochranu digitálních dat (*Long-term preservation* – LTP) systémů. Problematika dlouhodobé ochrany se neustále vyvíjí, stejně jako nástroje. Mnohé knihovny za posledních pět let zcela upustily od hybridní metody a všechny finance investují do digitalizace, mikrofilmy již nevytvářejí – např. NK Nizozemí, Britská knihovna a jiné. NK ČR se k tomuto trendu přidá v roce 2012, kdy bude v rámci projektu NDK spuštěna nová linka digitalizace, včetně přechodů na nové obrazové formáty i na nové formáty metadat.

K urychlení přijetí digitalizace jako odpovídajícího způsobu ochrany knihovních fondů přispělo i to, že se ve velké míře začaly objevovat tzv. digital-born dokumenty, které bylo nutno ochránit a uchovat pro budoucnost. Ochrana knihovních fondů se rozrostla o novou oblast, která byla zcela jiná, nijak nespojená s fyzickými dokumenty, a tuto oblast bylo nutno rychle začít řešit. Najednou tu nebyly jen zdigitalizované předlohy, ale i dokumenty, které neměly žádnou fyzickou předlohu ani kopii.

3.1.2 Digitalizace „pro zpřístupnění“

Zhruba od roku 2005 lze pozorovat obrat v konceptu digitalizace, z digitalizace „pro ochranu“ se stále více stává digitalizace „pro zpřístupnění“[12]. Digitalizace jako ochrana dokumentů se nevytrácí, ale daleko větší podporu má digitalizace jako metoda zlepšení zpřístupnění.[13] Digitalizují se tak i knihy, které nejsou nijak ohroženy ve své fyzické podstatě. Cílem je nabídnout víceméně vše v digitální podobě tak, aby si knihovny a paměťové instituce uchovaly svou podstatu i v nové době, kdy stále více a více uživatelů má pocit, že to, co není dostupné online, neexistuje – viz také [KAHLE, 2004, s. 31]. Markantní je to u dětí a mládeže, které se narodily již v době Internetu a svět před Internetem neznají. Od malička je tato skupina uživatelů zvyklá být online, návštěvě knihovny jako instituce a půjčování fyzických knih se brání. Jedná se o tzv. *digital natives*[14], dnes již více než deset let běžně používaný výraz - více např. viz [PRENSKY, 2001] nebo [PALFREY a GASSER, 2008]. Od roku 2003 se velmi důkladně řešily otázky, pro koho se vlastně digitalizuje, jaké jsou nároky uživatelů, jaké očekávají rozhranní a přes jaké technologie k digitálním knihovnám přistupují. Často se hledalo vyhranění nebo naopak přiblížení komerčním službám a webům, jako jsou Google a byly projekty digitalizace Yahoo a Microsoft apod. Tento směr je v současné době mírně upozaděn, ve prospěch dlouhodobé ochrany digitálních dat.

Digitalizace „pro zpřístupnění“ je spojena s digitálními knihovnami daleko více, než digitalizace „pro ochranu“. Digitální knihovna oproti běžné „kamenné“ knihovně má často uváděné výhody, mezi kterými jsou neustálá dostupnost dokumentů v podstatě odkudkoliv, zpřístupnění jednoho dokumentu takřka neomezenému počtu uživatelů v jediném okamžiku, široké možnosti vyhledávání, včetně plných textů a napojení dalších digitálních knihoven, šetření místa apod. Nevýhodou mohou být naopak vysoké náklady na HW, uložení, nutnost sledování autorského zákona, tvorby metadat; zastarávání formátů i technologií apod.

3.1.3 Projekty digitalizace, podpora Evropské unie

Projekty digitalizace vznikaly živelně v různých zemích a bylo jasné, že je potřeba z mnoha důvodů je koordinovat. V Evropské unii byl tento směr koordinace na úrovni členských států patrný od roku 2001, kdy byl za švédského předsednictví odsouhlasen tzv. *Lundský plán*. Principy a zásady ze zasedání expertů členských zemí ve Švédském Lundu byly přetvořeny do akčního plánu, kterým se nadále EU měla řídit a také řídila. Bylo zřejmé, že je nutné kontrolovat kvalitu digitalizace, určovat standardy a best-practices pro digitalizaci i tvorbu metadat, sledovat duplicity, podporovat interoperabilitu a sdílení zkušeností a znalostí. V plánu se ovšem neobjevují problémy související s řešením dlouhodobé ochrany digitálních dat, tj. výstupů vlastní digitalizace. Nešlo pouze o standardy, ale o souhru jednotlivých strategií členských zemí, jejich postupů a snah. Členské země (tehdy 15 zemí) odsouhlasily vznik stálého výboru expertů pro koordinaci. Skupina národních

[12] Dnes by šlo hovořit také o digitalizaci „pro čtení“. Cílovým zařízením nejen zdigitalizovaných dokumentů se totiž v posledních letech stávají elektronické čtečky a tablety, které jsou schopné si poradit jak s obrázky textu, tak s PDF a dalšími formáty.

[13] Digitalizace není první technologií, která pomohla ochraně fyzických dokumentů a zlepšila možnosti jejich zpřístupnění. Podobně mikrofilmy se po zdokonalení techniky začaly využívat jako materiál pro sdílení textů nezávisle na předloze. Mikrofilmy lze duplikovat a pak dále distribuovat. Z mikrofilmu je možné zpětně vytvořit i tištěnou stranu nebo kompletní dokument (např. pomocí tzv. copyflo machine).

[14] Lidé, kteří žili již v době před Internetem a v současnosti nové technologie využívají, se nazývají *digital immigrants*.

reprezentantů NRG (*National Representatives Group*) se věnovala národním strategiím pro digitalizaci, výměně informací apod. *Lundský plán* rozváděla více tzv. *Parmská charta*[15]. Jde o dokument schválený skupinou reprezentantů členských států EU při příležitosti 5. setkání skupiny NRG v Parmě, 19. 11. 2003. Charta obsahovala deset článků popisujících oblasti dalšího nutného rozvoje a také část popisující následující postup a cíle NRG skupiny. Články se týkaly podpory přístupnosti digitálního obsahu v Evropě, podpory kvality a standardů, ochrany autorských práv, interoperability, vícejazyčnému přístupu, hodnocení, a kooperace.

Tomuto tématu bylo věnováno mnoho konferencí a následných projektů – viz např. konference *Strategies for a european area of digital cultural resources: towards a continuum of digital heritage* konaná u příležitosti nizozemského předsednictví EU v roce 2004. První z velkých aktivit souvisejících s digitalizací byl *eContent* program (2001-2004), jehož cílem bylo prozkoumat vícejazyčná řešení přístupů k digitálnímu obsahu, databázím apod. Následný *eContent Plus* (2005-2008) byl rámcem, ze kterého byly financovány mnohé projekty digitalizace a související aktivity. Projekty *eContent Plus* byly z oblasti kontroly kvality obsahu, zkoumání potřeb uživatelů, vzdělávání. Jedním z projektů byla od roku 2002 MINERVA (*MInisterial NEtwoRk for Valorising Activities in digitisation*), která mj. podporovala *Lundský plán*, pořádala pravidelné schůzky expertů, byly tvořeny návody, postupy [Conference Day 1 & 2: a Brief Conference Report, 2004, s. 8]. *MINERVA* i pozdější *MINERVAPlus*, která zahrnovala do EU nově přistupující země, byly od počátku zaměřené na digitalizaci významných částí kulturního dědictví zemí zapojených v projektu [BAUEROVÁ, 2006a, s. 13]. Tyto projekty měly dvě roviny, jednu politickou, která znamenala podporu aktivitám digitalizace v konkrétních zemích, technická rovina zahrnovala sdílení znalostí, best practices apod.

Velkým úspěchem a podporou vývoje digitálních knihoven na evropském kontinentu byl projekt *DELOS*, který běžel mezi lety 2002-2006 a vyústil v projekt *DL.org*[16] podněcující interoperabilitu digitálních knihoven a tvorbu best practices. *DELOS* vytvořil významnou síť odborníků na digitální knihovny (*Network of Excellence*) a prováděl výzkum v oblasti architektury digitálních knihoven, personalizace i ochrany dat.

Digitální knihovny se měly stát spojovacím prvkem různých kultur EU. Komisařka pro Informační společnost a média Viviane Reding[17] situaci v oblasti digitálních knihoven komentuje slovy: *„Bez kolektivní paměti nejsme nic a nemůžeme ničeho dosáhnout. Definuje totiž naši identitu a můžeme ji používat pro vzdělávání, práci i zábavu.“* [EUROPEAN COMMISSION, 2005b]

Podpora řešení problémů digitalizace analogových dokumentů pro jejich uchování i pro lepší zpřístupnění v jednotlivých členských státech EU byla jednou ze tří hlavních částí evropské iniciativy *i2010: Digitální knihovny – Evropská informační společnost pro růst a zaměstnanost*. *i2010*, která byla představena na jaře 2005, vycházela z politiky Evropské komise vedené v rámci Lisabonské strategie, pro kterou měla představovat nový start [BAUEROVÁ, 2006b, s. 15]. Iniciativa se věnovala mimo podpory digitalizace také podpoře on-line zpřístupnění (projekty TEL, TEL-ME-MOR) a nově také podpoře dlouhodobé ochrany digitálního obsahu. Komisařka pro informační společnost a média Viviane Reding, nazvala tyto tři oblasti hlavními pilíři *i2010*. Za hlavní cíl dlouhodobé ochrany dat považuje iniciativa dlouhodobé a vlastně časem neomezené zpřístupnění zdigitalizovaných dokumentů. A to i navzdory velmi rychlému vývoj technologií SW i

[15] http://www.minervaeurope.org/structure/nrg/documents/charterparma.htm
[16] http://www.dlorg.eu/
[17] Ve funkci evropské komisařky pro informační společnost a média byla Viviane Reding od 22. listopadu 2004 do 9. února 2010.

HW, omezené životnosti datových nosičů a neexistenci národních strategií pro dlouhodobou ochranu digitálních dat [EUROPEAN COMMISSION, 2005a]. Jednou z odpovědí na tyto vyjmenované problémy dlouhodobé ochrany dat byl i sedmý rámcový program (FP7), kde uspěly projekty zaměřené na výzkum (CASPAR, PLANETS) a popularizaci problémů (*DigitalPreservationEurope*) dlouhodobé ochrany.

3.1.4 Ochrana digitalizovaných dat

Až do roku 2005 byl kladen hlavní důraz na zpřístupnění analogových dokumentů pomocí digitalizace. Otázka ochrany nebyla příliš brána v potaz a málokdo ji chtěl financovat. Ukázalo se ovšem, že problematiku dlouhodobé ochrany dat není možné ignorovat, protože její nutnost je velmi rychle viditelná. O zranitelnosti a problémech s digitálními daty se velmi dobře vědělo a s různými typy poškození, nekompatibilitou formátů apod. se instituce setkávaly. Proto po etapách „digitalizace pro ochranu", „digitalizace pro zpřístupnění" začala éra, kterou je možné nazvat „ochrana dat vzniklých digitalizací" – více viz kapitola 3.2. Průzkum provedený v roce 2005 Aliancí IFLA-CDNL pro bibliografické standardy – ICABS[18] pro nizozemskou Královskou knihovnu nazvaný *„Building Networks in Digital Preservation: Recent Developments in Digital Preservation in 15 National Libraries"* ukázal, že knihovny dosud nepřijaly jednotnou strategii pro dosažení dlouhodobé ochrany a přístupu k digitálním objektům, které přibývají do jejich sbírek. Některé z knihoven neměly ani žádnou strategii a to navzdory zřejmé hrozbě nebezpečí skrývajícího se v nedostatečném opatrování digitálních sbírek [VERHEUL, 2006]. Průzkum národních a lokálních archivů tak jen potvrdil výsledky, které vzešly již ze zprávy Hanse Hofmana a Maurizia Lunghi *„Enabling Persistent And Sustainable Digital Cultural Heritage in Europe"* z roku 2004 [HOFMAN a LUNGHI, 2004].

Současná situace je taková, že dlouhodobá ochrana digitálních dat se stala nedílnou součástí projektů digitalizace. Udržitelnost, ochrana a uložení pro budoucí použití digitálních objektů je vyžadováno u všech EU projektů. Jako čtvrtou ze čtyř fází digitalizace samotné uvádí dlouhodobou ochranu digitálních dat mj. [BÜLOW a AHMON, 2011, s. 11]. Přínos, který digitalizace má v podobě online zpřístupnění, bude přínosem pouze tak dlouho, dokud budou digitální objekty vyhledatelné, zobrazitelné a pochopitelné. Aby tomu tak bylo např. i za sto let, je hlavním úkolem právě dlouhodobé ochrany digitálních dat.

3.1.5 Digitalizace v NK ČR – stručný přehled

Počátkem 90. let 20. století byla jednou z hlavních priorit NK ČR automatizace knihovnických procesů a do této oblasti plynula většina financí určených na informační technologie. Digitalizace mezi prioritami nebyla. Již první léta po roce 1990 přinesla první viditelné nadšení z možností, které může digitalizace přinést ochraně fyzických fondů díky novým možnostem zpřístupnění. UNESCO zahajovalo program *Paměť světa*, ve kterém počítalo s využitím CD disků pro zpřístupnění [KNOLL, 2010a, s. 21]. NK ČR se této výzvy chopila a vznikly tak první digitalizované obrazy a první CD pro UNESCO *Paměť světa*. Později další CD s již kompletními digitálními verzemi fyzických předloh. V roce 1995 se NK ČR ve spolupráci s firmou AIP (Albertina Icome Praha) rozhoduje postavit první digitalizační pracoviště pro digitalizaci historických dokumentů a o rok později byla zahájena rutinní digitalizace rukopisů a starých tisků [KNOLL, 2010a, s. 22-24]. Až v

[18] http://www.ifla.org/VI/7/icabs.htm

roce 1999 se začalo s digitalizací novodobých dokumentů, konkrétně periodik a to z mikrofilmů.[19] Periodika z 19. a 20. století nebyla zpočátku prioritou, důležitější se zdály být rukopisy a staré tisky.[20] Přitom periodika jsou velmi žádána čtenáři a badateli, často jsou dochována pouze v jediném exempláři a namáhána tak častým půjčováním.

V roce 2000 byl zahájen grantový program Ministerstva kultury ČR pod názvem VISK, kde má digitalizace své místo v podprogramech VISK4, VISK6 a VISK7 – viz heslo VISK v kapitole 2.1. Ve stejné době se začíná NK ČR angažovat v evropských projektech, např. MASTER. Po roce 2000 lze pozorovat větší aktivitu jak na poli samotné digitalizace, tak na poli metadat. Začínají se rozlišovat jednotlivé typy dokumentů a vznikají specializované aplikace zpřístupnění (Manuscriptorium a Kramerius). Metadatový popis historických a novodobých dokumentů se také vydává odlišnými cestami a začíná existovat na dvou platformách (MASTER TEI vs. DTD periodika a monografie) – více viz kapitola 5.

Obrovským impulsem pro rozvoj digitalizace novodobých periodik a rozvoj infrastruktury byly povodně v roce 2002, které poničily mnoho fondů různých knihoven. Jako možnost záchrany a zpřístupnění takto poničených fondů se nabízela právě digitalizace. K digitalizaci i mikrofilmování byl využit již fungující program VISK7 a od roku 2003 i open source aplikace Kramerius na zpřístupnění novodobých digitálních dokumentů na Internetu, která tvořila stejnojmennou digitální knihovnu.[21] S digitalizací novodobých monografií se začalo v projektu *Kramerius* až v roce 2006. Úspěchem NK ČR bylo v roce 2005 ocenění JIKJI v jihokorejském Soulu, za celosvětový přínos v oblasti digitalizace, převážně díky digitalizaci historických fondů.

Skutečná situace v digitalizaci v NK ČR a v ČR obecně ovšem byla tristní. Rozpočty na digitalizaci novodobých fondů byly snižovány, což způsobilo pomalý postup digitalizace. V koncepci rozvoje knihoven v ČR na léta 2004 až 2010 je digitalizace zmíněna v bodu 24, který má název *Pokračovat v digitalizaci vybraných částí knihovních fondů jako součásti kulturního dědictví a zpřístupnit je veřejnosti* [ČESKO. MINISTERSTVO KULTURY, 2004, s. 32]. Digitalizace se obecně prolíná celým dokumentem, který klade důraz na spolupráci knihoven, na větším zpřístupnění sbírek uživatelům klasickou ale i elektronickou formou. Hlavní podpora digitalizace měla být realizována pomocí stávajících programů VISK 4, 6 a 7 s vidinou podstatného zrychlení digitalizace, možná i se zavedením digitalizace masové. Jak se ale ukázalo v následujících letech, financování těchto programů bylo stále více a více kráceno. Již z původně plánovaných zdrojů na financování projektu VISK mezi lety 2000-2005 ve výši 777,6 milionů Kč, byly skutečně dostupné prostředky pouze ve výši 288,6 milionů Kč, tj. pouhých 37% [Koncepce trvalého uchování…, 2005, s. 17]. V letech 2005-2010 byla situace obdobná, ke konci tohoto období se financí dostávalo stále méně a méně. K odpovídajícímu naplnění koncepce v tomto ohledu tedy nedošlo a digitalizace v NK ČR i obecně v českém prostředí fungovala určitou setrvačností na zastaralých technologiích a v menších objemech, než bylo plánováno a potřeba. I tak byly programy VISK 6 a VISK 7 tahouny veškerého snažení v oblasti digitalizace po celé první desetiletí 21. století.

[19] Projekt Digitalizace mikromédií 1997–1999 a Digitální knihovna: produkce, ochrana a zpřístupnění digitálních dokumentů 1999–2003.

[20] U periodik se předpokládalo mnoho problémů, které bude jejich zpracování přinášet, proto se s nimi nezačalo dříve (např. velikost formátu jednotlivých výtisků, obrovská rozsáhlost jednotlivých titulů (desítky ročníků, tisíce čísel) apod.).

[21] Díky tomu je dodnes matoucí, když někdo mluví o Krameriovi bez dalšího upřesnění, nemusí být jasné, zda se jedná o projekt, systém nebo digitální knihovnu.

Vše se začalo zlepšovat až díky externím zdrojům financování. Velkou vzpruhou byly od roku 2007 tzv. *Norské fondy*, díky kterým se začaly ve větší míře digitalizovat monografie z 19. století. V letech 2007-2009 bylo zdigitalizováno, metadaty opatřeno a zpřístupněno 2,4 milionů stran.[22] Projekt se jmenoval *Záchrana neperiodických bohemikálních dokumentů 19. stol. ohrožených degradací papíru*, dostal finanční dotaci ve výši 999 960 Euro [ČESKO. MINISTERSTVO FINANCÍ, 2011]. Digitalizace probíhala skenováním nasnímaných mikrofilmů a také hybridní metodou.
Konec prvního desetiletí 21. století je ve znamení zlepšování zpřístupnění[23] a snahy začít s masovou digitalizací. Za obrovský úspěch lze také považovat celkem 10 milionů naskenovaných stran novodobých periodik a monografií k roku 2011 v projektu VISK7, i když možnosti zvyšovat produkci digitalizace zůstávaly díky nedostatečnému financování nevyužité.

3.1.5.1 Masová digitalizace a projekt Národní digitální knihovna

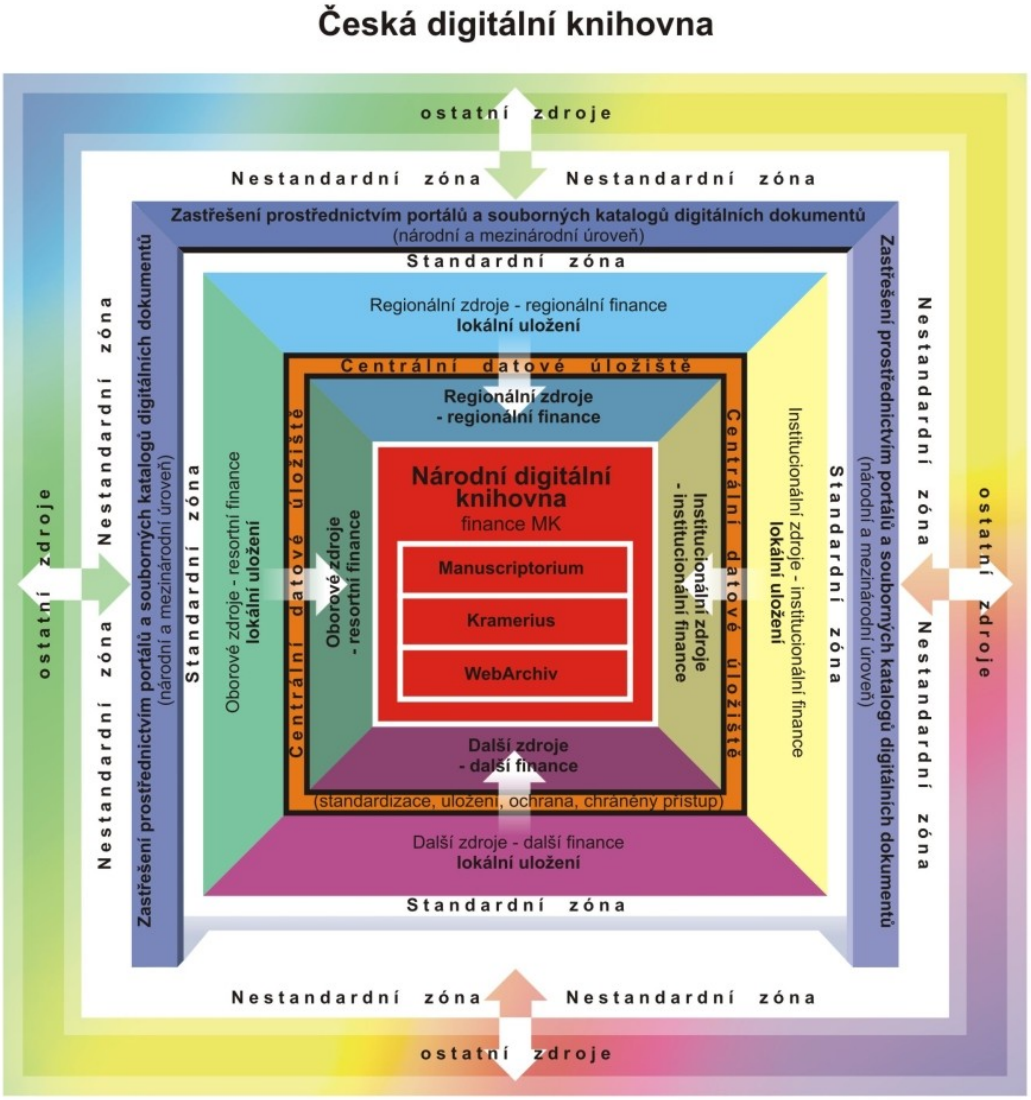
O masové digitalizaci, která již několik let probíhá v západní Evropě a vyspělém světě[24], se začalo v NK ČR přemýšlet jako o možnosti pro další urychlení digitalizace až v roce 2007, kdy odstartovala jednání o spolupráci s firmou Google v projektu *Google Books*[25]. Neuvažovalo se ještě o vlastních robotických skenerech a vlastní masové digitalizaci. Společnost Google si vyžádala rozbor fondů NK ČR z let 1620-1900 s tím, že zvažuje masovou digitalizaci obdobných fondů v Evropě. Nicméně vzhledem k aktivitám Google mimo Evropu jednání usnulo [KNOLL, POLIŠENSKÝ a UHLÍŘ, 2007, s. 3]. K dohodě došlo až před vánoci 2010, kdy se NK ČR stala jednou z evropských knihoven, kterým Google zdigitalizuje některé z jejích svazků výměnou za přípravu metadat a vlastní data, kterých bude vlastníkem. Půjde o 200.000 svazků vydaných do konce 19. století, včetně dokumentů ze Slovanské knihovny.
K úvahám o nákupu vlastních robotických skenerů došlo v roce 2008, kdy se začal formovat projekt NDK, který má masovou digitalizaci jako jeden ze svých tří hlavních cílů. Robotické skenery, které jsou schopny automaticky digitalizovat 1000 stran za hodinu[26], budou do NK ČR a do Moravské zemské knihovny (MZK) nakoupeny v roce 2012. Oproti ostatním vyspělým národním knihovnám jde o několikaleté zpoždění s nasazením této technologie a to i přesto, že již v roce 2005 o robotických skenerech Kirtas a 4DigitalBooks mluvil Ivo Iossieger.[27] Pracovníci AIP Beroun tehdy testovali 4DigitalBooks skenery na historických novinách z NK ČR [PSOHLAVEC, 2006, s. 39]. Výsledky byly nadějné, ovšem k jejich nákupu do NK ČR nebo jinam do českých paměťových institucí nedošlo. V rámci projektu NDK, ve kterém budou v NK ČR poprvé nasazeny

[22] Vzniklo i 1,2 milionu matričních negativů plus stejný počet negativů archivních.
[23] Vytvoření virtuálního badatelského prostředí pro zpřístupnění a ochranu digitálních dokumentů 2004–2010.
[24] Doporučení Evropské komise členským státům ze 24. srpna 2006 zavazuje státy, aby vytvořily pracoviště masové digitalizace a provozovaly [KNOLL, POLIŠENSKÝ a UHLÍŘ, 2006, s. 19].
[25] V prvních letech nového tisíciletí se začaly objevovat velké projekty kooperace mezi paměťovými institucemi a komerčními firmami. V knihovnách to je od roku 2004 známý *Google Books* (dříve *Google Library Project*), kterého se mj. účastní od roku 2010 i NK ČR. Byl to také obdobný projekt společnosti Microsoft – *Microsoft Live Books*, který byl ukončen v roce 2008.
[26] Je nutno si dát pozor na to, že robotické skenery nedosahují hodnot udávaných výrobci. Ti někdy uvádí i více než 2000 stran za hodinu. Ze zkušeností a z debat s kolegy z různých evropských zemí (Slovensko, Nizozemí, Rakousko apod.) ovšem vyplývá, že reálná rychlost je na hranici 1000 stran za hodinu, a to ještě ve velmi optimálních podmínkách a s ideálním dokumentem.
[27] Firmy jako Treventus, Quidenus aj. začaly vyvíjet a nabízet své robotické skenery později, komerčně dostupné jsou od roku 2008.

robotické skenery, vzniknou v roce 2012 dvě pracoviště digitalizace, jedno v Klementinu, druhé v Moravské zemské knihovně, která je partnerem projektu NDK. Denní maximální kapacita by měla být zhruba 80 tisíc stran.

Celý proces digitalizace v NK ČR během doby nalézal záchytné body a své opodstatnění ve strategických dokumentech, jakou byly např. *Státní informační a komunikační politika* [ČESKO. MINISTERSTVO KULTURY, 2004], *Koncepce rozvoje knihoven v České republice na léta 2004 až 2010* [ČESKO, 2004] a konečně *Koncepce trvalého uchování knihovních sbírek tradičních a elektronických dokumentů v knihovnách ČR do roku 2010* [Koncepce trvalého uchování…, 2005], kterou vytvořila NK ČR ve spolupráci s Ministerstvem kultury ČR. *Koncepce trvalého uchování knihovních sbírek tradičních a elektronických dokumentů v knihovnách ČR do roku 2010* rozpracovávala některé dílčí body *Koncepce rozvoje knihoven na léta 2004 až 2010* a poprvé se objevil pojem *Národní digitální knihovna*, která měla být jádrem širšího pojetí v podobě *České digitální knihovny*. Koncepce se věnovala nejen digitalizaci, kterou měla urychlit, ale poprvé také problematice uchování dokumentů, které má NK ČR za povinnost vyplývající z knihovnického zákona, a která nebyla do té doby uspokojivě formulována a ani řešena – viz kapitola 5.3.3. Za základ *Národní digitální knihovny* jsou od počátku považovány tři největší projekty NK ČR a to *Kramerius*, *WebArchiv*[28] a *Manuscriptorium*, jejichž obsah tvoří základ národního kulturního dědictví. *Česká digitální knihovna* je tvořena dalšími digitálními dokumenty, které se nekvalifikují pro uložení do NDK a mohou být regionálního, oborového a jiného charakteru [STOKLASOVÁ, 2006, s. 110] – viz Obrázek 1. Za tato data a digitální knihovny nese zodpovědnost jejich majitel, zřizovatel instituce (ministerstva) apod. Za obsah NDK ovšem nese zodpovědnost NK ČR a tedy přeneseně Ministerstvo kultury ČR.

[28] http://www.webarchiv.cz/



Obrázek 1 – Česká digitální knihovna (autor PhDr. Bohdana Stoklasová, výtvarný návrh Martin Zhouf).

Koncept *Národní digitální knihovny* získal později financování ze strukturálních fondů EU ve výši 300 milionů Kč[29] a vznikl z něj projekt *Vytvoření Národní digitální knihovny*, který běží v NK ČR a MZK od roku 2008 a má být ukončen v roce 2014. Tento projekt má za cíl vyřešit nejbolavější místa českého snažení v oblasti digitalizace, kterými jsou nedostatek financí na zavedení masové digitalizace, neexistující dlouhodobá ochrana digitálních dat a snaha o vytvoření jednotného uživatelského rozhraní pro přístup ke všem možným zdrojům informací dostupných v dnešních moderních knihovnách. Do konce roku 2014, tedy do ukončení financování z EU, má být zdigitalizováno 26 milionů stran periodik a monografií z 19. a 20. století. Takto vzniklá data mají být bezpečně uložena v LTP systému, který zajistí jejich dlouhodobou ochranu a zpřístupnění v budoucnu, včetně ochrany pro data vzniklá digitalizací do roku 2010. Ta jsou nyní uložena v archivu, který zajišťuje pouze uchování bitstreamu. Díky získaným zkušenostem a nakoupeným

[29] Původně byl projekt plánován na 700 milionů Kč, po vytvoření první studie proveditelnosti v roce 2009 byl rozpočet krácen, musela se tak měnit i studie a propozice celého projektu.

systémům se bude v digitalizaci pokračovat. Projekt je nastaven tak, aby bylo možné digitalizovat do roku 2019 v rámci udržitelnosti investice. Do té doby by mělo být hotovo 50 milionů naskenovaných stran, což odpovídá asi 300.000 svazků, a podstatná část českého kulturního dědictví uloženého v knihách a časopisech by tak měla být zachráněna. Rychlostí digitalizace, kterou se postupovalo do roku 2011, by se takový počet stránek digitalizoval více než 100 let.

3.2 Dlouhodobá ochrana digitálních dat – hlavní problémy, možnosti řešení a vývoj pohledu na ni

V předchozích kapitolách o digitalizaci je popsán problém, který digitální objekty provází. Je spojen se zastaráváním SW a HW, s formáty objektů, s možnou ztrátou kontextu a celkovou zranitelností digitálních objektů. Netýká se ovšem pouze zdigitalizovaných dokumentů, ale všech digitálních objektů, ať již vznikly jakkoliv. Již od vynálezu prvního programovatelného počítače v roce 1936 lze sledovat postupný, ovšem zrychlující proces přechodu od informací držených v analogové podobě k digitálním médiím. Tento přerod je možný díky tomu, jak elektronická informace ulehčuje sdílení, kopírování i vlastní uložení. V posledních letech dochází k velkému nárůstu množství digitálních dat ve všech oblastech našeho života. Již dávno nejsou velké objemy dat spojeny pouze s IT průmyslem nebo vědou, ale se všemi oblastmi všedního života. Dennodenně každý z nás vytváří nějaká digitální data, ať už v zaměstnání nebo třeba na dovolené. S tímto růstem se pojí základní otázka, jak tato data ochránit a zajistit jejich použitelnost a důvěryhodnost v budoucnu. Otázky se stávají palčivějšími v okamžiku, kdy se jedná o velké objemy dat, které jsou například výsledky několikaletého úsilí, např. vědeckých výzkumů, vývoje nebo skenování fyzických předloh. UNESCO, v obsáhlém dokumentu *Guidelines for the Preservation of Digital Heritage* z roku 2003 [WEBB, 2003, s. 29-30] vyjmenovává všechny typy digitálních objektů, které je potřeba považovat za kulturní dědictví a tedy ochraňovat. V seznamu jsou např. elektronické publikace na webu, optických discích; nepublikovaná literatura (akademická práce, e-printy); vládní a firemní dokumenty; databáze; výukové dokumenty; SW nástroje; elektronické rukopisy (osobní korespondence); zábavné materiály (filmy, hudba apod.); zdigitalizované dokumenty aj. Seznam by dnes mohl být delší.

Ve své první zprávě *Digital Universe* z roku 2007, kterou každoročně vydává společnost *International Data Corporation* (IDC), je odhadována velikost existujícího digitálního prostředí (vesmíru) v počtu bitů (tedy jedniček a nul) na 1,288 x 10^{18}, což činí 161 bilionů GB (exabytů). V roce 2008 zpráva uvádí 281 exabytů a uvádí odhad o desetinásobném růstu do roku 2011. V roce 2009 zpráva IDC odhaduje celkovou množinu digitálních dat na 481 exabytů a toto množství se má zdvojnásobit každých 18 měsíců [BONIN, 2009, s. 3]. V roce 2010 byla prolomena hranice zettabytu a v roce 2011 měla celková množina digitálních informací velikost 1,8 zettabytů (1,8 trilionu GB). Pouze třetina těchto informací má nějaký typ zálohy nebo zabezpečení, pouze polovina informací, které by měly být zabezpečeny nějaké zabezpečení má [GANTZ a REINSEL, 2011, s. 1]. Ironií je, že dokážeme uložit kvanta dat na velmi malých fyzických médiích, ale tato uložená data nedokážeme stále efektivně ochránit pro budoucí využití. Ještě nikdy nebyla autenticita, integrita, kompletnost a celistvost dokumentů tak důležitá, jako v případě dokumentů v digitální podobě. Mnoho organizací, které digitální informace již delší dobu využívají, se nyní dostává do situace, kdy musejí, chtě, nechtě přistoupit k nějakým „archivním“ opatřením.

Paměťové instituce, a nejen ony, mají za povinnost chránit a zpřístupňovat uložené dokumenty. Ochránit a zachovat digitální data v dlouhodobém horizontu je ovšem daleko těžší než např.

v případě papíru. Ve světě tradičních dokumentů bylo dlouhodobé uchování výrazně jednodušší. Dokumenty ze své fyzické podstaty vydržely desítky i stovky let. Degradace papírových nosičů informací je pomalá a snadno zjistitelná, ztráty v digitálním světě jsou naopak rychlé, nevratné a ne vždy snadno a včas zachytitelné [STOKLASOVÁ a HUTAŘ, 2007, s. 89]. Pro papírový dokument není potřeba žádná technologie k jeho čtení nebo prohlížení. Právě HW a SW nutný pro zobrazení digitálních objektů zastarává během let (někdy i měsíců). Především ale na rozdíl od fyzických dokumentů dokumenty digitální neexistují v hmatatelné podobě. Před nástupem Internetu panoval názor, že ochrana digitálních dokumentů spočívá v ochraně nosiče/média, které je nese. Internet ukázal, že to byl chybný kalkul. Dnes nejsou digitální objekty vázány na žádný nosič a jejich dlouhodobá ochrana spočívá v udržení jejich vlastností, integrity, autenticity během času.
Neochraňujeme nosič jako v případě fyzických dokumentů, ale informaci samotnou, tedy data, bitstream, který je někde uložen a přenášen po sítích nebo jiných fyzických médiích. K tomu, abychom objekt mohli ochránit, spravovat a udržovat jej, měnit jeho formát vhodný pro nový SW apod. potřebujeme metadata. Digitální objekt bez kontextu a doprovodných údajů v podobě metadat je nepoužitelný. Abdelaziz Abid v jednom ze svých článků připomíná, že ochrana digitálních dat je nikdy nekončící proces: "*Pro ochranu analogových dokumentů většinou stačí uložení v optimálních podmínkách, dostatečná kontrola fyzického stavu a minimální využívání. U digitálních dokumentů je situace podstatně složitější. Jejich ochranu lze přirovnat k udržování ohně – je nutné se mu věnovat neustále, udržovat ho a kontrolovat. Jinak zhasne a nenávratně zmizí. Při správné péči ale může být věčný.*" [ABID, 2007, s. 7]
Digitální informace jsou závislé na technologii, která je uloží, dekóduje a zobrazí. I díky tomu je elektronické dokumenty velmi jednoduché měnit a upravovat, aniž by to bylo patrné (při záměrné změně). Změny mohou nastat také nechtěné, např. během přenosu, opět často aniž by si toho někdo všiml. Z těchto důvodů je nutné udržovat údaje v podobě metadat o provenienci, o změnách a je nutné zajistit i určitou úroveň bezpečnosti proti neoprávněným přístupům nebo nechtěným změnám, poruchám. Flexibilita digitálních informací je jejich výhodou, ale také problémem. Digitální objekt není v podstatě nikdy ve finální podobě, jak ji známe z prostředí papírového, vždy se může změnit a mít tak novou verzi.
Uložení tradičních dokumentů bylo statické, a třebaže si žádalo určité výlohy, představovaly jen nepatrný zlomek ekonomické náročnosti na dynamickou archivaci digitálních objektů. Náklady na uchování digitálních informací jsou velmi vysoké, mluvíme-li například o pořízení odpovídajícího LTP systému, jehož cena se pohybuje v řádech desítek milionů korun. Pominout samozřejmě nelze ani běžné provozní a vývojářské náklady, které si takový systém vyžádá. Jen náklady na údržbu se většinou uvádějí ve výši okolo 10% ceny zařízení/systému za rok. S financemi souvisí otázka výběru digitálních dokumentů k dlouhodobé ochraně. Představy o tom, že všechny digitální objekty konkrétní instituce budou dlouhodobě ochraňovány, nejsou zcela reálné. Často je problémem vůbec odhadnout co je kulturní dědictví s přetrvávající hodnotou. Instituce většinou přistupují k rozdělení dokumentů do skupin podle významnosti, od kterého se odvíjí úroveň poskytnuté dlouhodobé ochrany (např. Národní knihovna Nizozemí). Ne pro všechny dokumenty je nutné vytvářet maximálně obsažná ochranná metadata, nebo s nimi provádět ochranné procesy (např. uživatelské kopie, duplicity).

Dlouhodobá ochrana digitálních dat je nyní v paměťových institucích otázkou velmi aktuální, která zastiňuje ostatní problémy. Problematika přesahuje jednotlivé země, jak akcentovala již v roce

2003 charta UNESCO o „O ochraně digitálního dědictví"[30] – více viz [ABID, 2007]. Jak ve svém článku *The Digital Dark Ages? Challenges In The Preservation Of Electronic Information* napsal Terry Kuny: *„Čím déle žijeme v elektronické éře digitálních dokumentů, tím více si musíme uvědomit, že … se dostáváme do období, kdy mnohé z toho co dnes víme, co je digitálně kódováno, bude navždy ztraceno. Žijeme na počátku období digitálního temna a je na knihovnících a knihovnách, stejně jako na mniších v dávných klášterech, aby udrželi tradici vedoucí k zachování dokumentů publikovaných v současnosti.*" [KUNY, 1998] Dlouhodobá ochrana digitálních dat není luxus. Zajištění odpovídající ochrany pro digitální objekty musí být součástí instituce, jejího statutu a denní náplně stejně jako je ochrana fyzických dokumentů před vlhkostí, ohněm apod.

3.2.1 Problémy spojené s digitálními objekty

Digitální objekty ze své podstaty jsou velmi křehké a náchylné k poškození. Jsou závislé na technologiích, což sebou přináší potíže. Níže uvedu několik oblastí, kde digitální objekty mají slabiny, a které by dlouhodobá ochrana dat měla brát v potaz a měla by je řešit. Velmi pěkný obecný úvod do problému zastarávání HW i SW, problémů budoucího zpřístupnění, správy dokumentů a vůbec uchování digitálních dat napsal Seamus Ross v eseji *Changing Trains at Wigan: Digital Preservation and the Future of Scholarship* [ROSS, 2000].

Problémy digitálních objektů mohou být fyzického nebo logického charakteru. Fyzickým problémem je stárnutí nebo poškození nosičů dat (opotřebovanost, špatné podmínky uložení, porucha infrastruktury, přírodní pohroma), případně zastarávání HW.

Problémy logického charakteru jsou komplikovanější. Jejich předcházení a řešení předpokládá i změnu bitstreamu, vytváření doplňujících údajů v podobě metadat, dokumentaci apod. Archivovaná informace v podobě digitálního objektu musí být použitelná uživateli, kteří jsou vzdáleni v čase, prostoru a nemají podporu producenta té informace. Může se stát, že producent informace již neexistuje a nemůže tak odpovědět na žádné otázky, které by vedly k objasnění jejího smyslu apod. Jediné, co máme k dispozici, je digitální objekt a jeho metadata (v ideálním případě). Také SW, na kterém byla informace (digitální objekt) vytvořena, již nemusí být podporován. Budoucí komunita může používat naprosto odlišné pracovní prostředí, ve kterém nebude možno použít původní podobu digitálního objektu. Může se tedy stát, že informace zaznamenané tímto SW jsou zcela nedostupné a nepoužitelné. Nikdo nezná kódování pro jejich využití a dokumentace není k dispozici. Předpokladem k tomu, aby se výše uvedené nestalo, jsou metadata, která udržují veškeré nutné údaje. Logická ochrana na jejich základě umožní migrace dat a dokumentace procesů, které vedou ke zpřístupnění i v budoucích technologických prostředích.

Podobnou komplikací je, že uživatelská komunita se bude během doby měnit. Budoucí komunita, i když může být cílovou komunitou, pro kterou digitální objekt vznikl, nemusí znát kontextové informace u konkrétních informací (vznik, SW, HW, účel vzniku), pokud tyto údaje nebudou v metadatech. Archivované objekty mohou být dobře ochráněny, ale pokud je nelze identifikovat, nemají dostačující popis, uživatelé je nenajdou, případně je nechápou nebo nevědí v jakém SW je otevřít. Všechny procesy dlouhodobé ochrany potřebují metadata. Výzvou je tato metadata zajistit v dostatečné míře a ideálně automatickou cestou co nejblíže okamžiku, kdy digitální objekt vzniká.

[30] http://unesdoc.unesco.org/images/0013/001331/133171e.pdf#page=80

3.2.2 Definice dlouhodobé ochrany digitálních dat

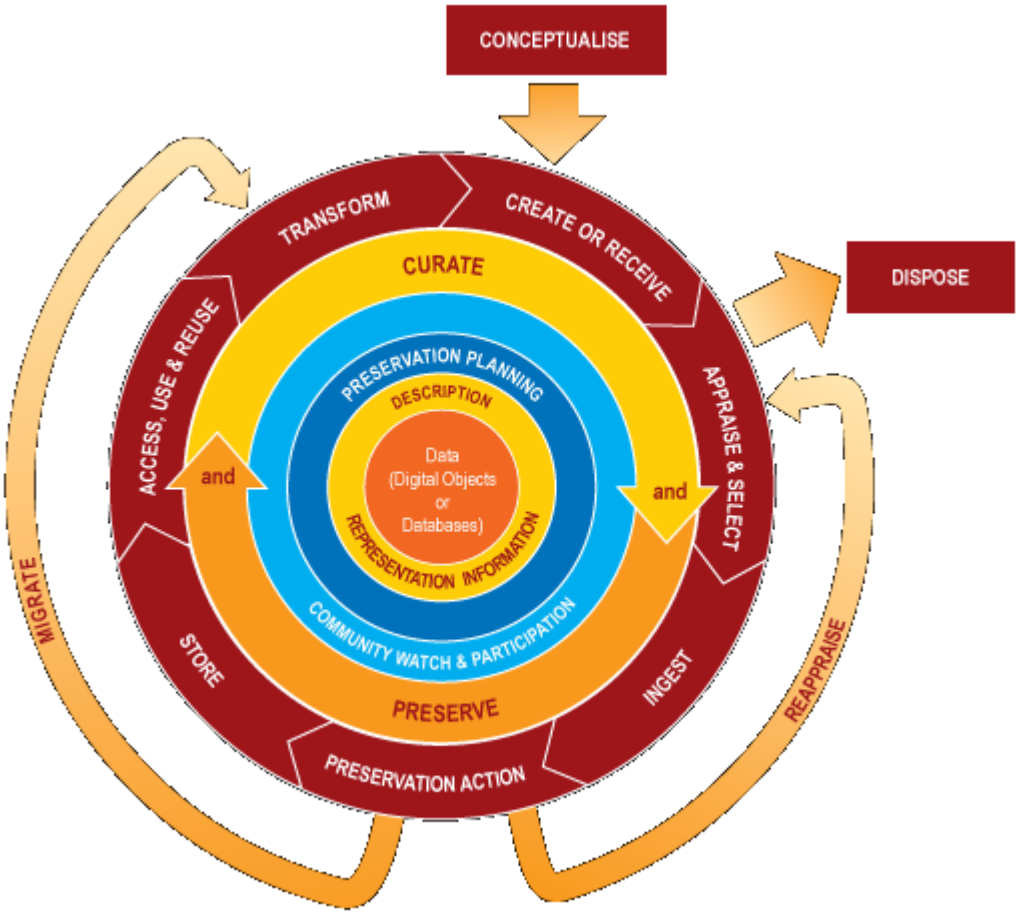
Dlouhodobá ochrana digitálních dat má ze své podstaty mnoho definic, které se liší úhlem pohledu na problematiku. Jedna z mnoha definic „*digital preservation*“, jak zní anglický výraz, vznikla v roce 2007 v rámci pracovní skupiny Americké asociace knihoven, která publikovala tři definice, krátkou, středně dlouhou a dlouhou. Středně dlouhá zní: *„Dlouhodobá ochrana digitálních dat spočívá v kombinaci plánů, strategií a opatření pro zajištění přístupu k reformátovanému a digitálně vzniklému digitálnímu obsahu bez ohledu na problémy spojené s nestálostí médií a technologickými změnami. Cílem dlouhodobé ochrany digitálních dat je přesné zobrazení autentického obsahu v jakémkoliv časovém horizontu od jeho vzniku.“* [ASSOCIATION FOR LIBRARY COLLECTIONS & TECHNICAL SERVICES, 2007]
Podobně H. M. Gladney [GLADNEY, 2007, s. 270] říká, že dlouhodobá ochrana digitálních dat je: *„organizované opatření k zajištění dlouhodobé použitelnosti digitálních objektů; zásadní je, že digitální objekty nebudou nikdy ztraceny nebo poškozeny, jsou důvěryhodné, je možné je vždy najít, rozumět jim a to i přes problémy zastarávání technologií.“*
Poměrně široká definice *Digital Preservation Coalition* (DPC) říká, že dlouhodobá ochrana digitálních dat jsou vlastně: *„všechny aktivity nutné k zajištění přístupnosti k digitálním materiálům i přes obtíže způsobené selháním médií nebo změnou technologií.*“ [DIGITAL PRESERVATION COALITION, 2009c] Z pohledu jednotlivých kroků dlouhodobou ochranu specifikuje Seamus Ross: *„Dlouhodobá ochrana digitálních dat má za cíl zajistit, že uživatelé v budoucnu budou moci digitální informace nalézt, získat, zobrazit, manipulovat s nimi, interpretovat je a to vše přes konstantní změny technologií. Celý proces zahrnuje konzervaci, obnovování, výběr, mazání, vylepšování, updatování a přidávání kontextu.*“ [ROSS, et al., 2009, s. 5]

Obecně tedy lze říci, že jde o kombinaci technických, právních a organizačních výzev a následných opatření, které by při jejich zvládnutí měly vést k bezpečnému uložení a především k možnosti *zpřístupnění* obsahů digitálních objektů i jejich kontextu v budoucnu. Ve většině projektů a plánování pro dlouhodobou ochranu digitálních dat není doba ochrany přesně specifikována, ale málokdo si myslí, že je dnes možné předpokládat a plánovat ochranu „navždy“. Cíle dlouhodobé ochrany jsou v realitě střednědobé s tím, že se stále průběžně upravují. Díky tomu musí být procesy ochrany kontinuální a měly by jít dále a dále do budoucnosti. Pokud se sami sebe zeptáme, jaký časový úsek označuje slovo „dlouhodobá“, můžeme si odpovědět např. tím, že ochrana má trvat potud, pokud obsah dokumentů někoho zajímá. Pakliže nezajímá nikoho, není nutné dokumenty chránit.

Dlouhodobá ochrana je součástí většího konceptu *digital curation*[31] (viz např. [HARVEY, 2010]). Podobně i [LAVOIE a DEMPSEY, 2004]: *„Dlouhodobá ochrana digitálních dat není izolovaný proces, ale komponenta širšího rámce vzájemně propojených služeb, strategií, vydavatelů, které dohromady tvoří digitální informační prostředí*“.
Digital curation se zabývá celým životním cyklem digitálního dokumentu a směřuje k jeho dlouhodobému uchování a použitelnosti. *Digital curation* začíná ještě před vznikem digitálního dokumentu, kdy se rozhoduje o jeho formátu, způsobu vzniku, použití vhodných metadat.

[31] Digital curation a digital preservation nejsou synonyma.

Pokračuje jeho uchováním, správou, případným doplňováním a změnami, kontrolou integrity a autenticity. To vše, aby byl použitelný a zpřístupnitelný budoucím uživatelům. Kontrolou integrity, zachováním autenticity, zpřístupnitelnosti a použitelnosti se dle Harveyho zabývá právě dlouhodobá ochrana digitálních dat [HARVEY, 2010, s. 55]. Viz Obrázek 2 znázorňující schéma modelu *DCC Digital Curation* životního cyklu, kde *digital preservation* (*preserve*) je pouze součástí celku.



Obrázek 2 – DCC Curation Lifecycle Mode [DIGITAL CURATION CENTRE, 2010].

3.2.2.1 Aktivní (logická) a pasivní dlouhodobá ochrana digitálních dat

Je nutno říci, že v ČR i ve světě stále panuje nepochopení, resp. nerozlišování mezi tzv. pasivní ochranou a aktivní logickou ochranou digitálních dat. Přitom jde o zásadní rozdíl. Pasivní ochrana je běžná a instituce ji provádějí. Spočívá v pouhé ochraně bitstreamu, tedy vytváření záloh, provozování více lokací apod. Ochrana bitstreamu ovšem neřeší problémy se zastaráváním SW, HW a formátů vlastních digitálních objektů, ani problémy s autenticitou. Pasivní ochrana, tedy uchování bitů, je pouze předpoklad, případně první krok, ochrany logické.

Naproti tomu logická dlouhodobá ochrana digitálních dat spočívá v kontrole a procesech prováděných během životního cyklu digitálního objektu tak, aby byla zajištěna jeho trvalá použitelnost. Použitelnost je obecný pojem pro vyhledatelnost, zobrazitelnost, pochopitelnost a autentičnost. Digitální objekty v archivu musí být srozumitelné a technicky použitelné v každé fázi svého životního cyklu. Archivní objekt není navždy neměnná entita uložená na pásce nebo harddisku [FOJTŮ, HUTAŘ a MELICHAR, 2011, s. 74-75]. Naopak, k zajištění trvalé použitelnosti

musí být dokumenty v archivu stále „živé“, reflektovat změny v globálním technickém prostředí a reagovat na změny, které vyžaduje administrativa dokumentů. Pro objekty musí být vytvořena a spravována odpovídající metadata (o vlastnostech, kontextu aj.), která se neustále doplňují a obohacují. Každá událost či změna archivního objektu má být zaznamenána v metadatech. Digitální objekty musí být neustále monitorovány z hlediska integrity a signifikantních vlastností. S tím vším souvisí neustálá kontrola organizace/instituce, která ochranu zajišťuje, pomocí auditů a jiných podobných mechanismů tak, aby organizace měla např. finance na ochranu, mandát apod. Logická ochrana digitálních dat využívá všech těchto údajů v podobě metadat k tomu, aby upozornila na hrozící problém a bylo možné kdykoliv provést ochranné akce (*preservation actions*), nejčastěji migraci nebo emulaci tak, aby byl zajištěn přístup k dokumentům i v budoucnu. K tématu viz také [SINCLAIR a BERNSTEIN, 2010]. Mnoho institucí považuje dlouhodobou ochranu za pouhou zálohu dat, nevidí tu aktivní část, procesy, které vedou k budoucímu využití digitálních objektů v novém technologickém prostředí.

3.2.3 Dlouhodobá ochrana digitálních dat a různé typy institucí

Paměťové instituce se posledních deset let snaží o zachycení, uložení a logickou dlouhodobou ochranu digitálních dokumentů. Až donedávna nebylo možno říci, že by byly příliš úspěšné. Reálně stále hrozí ztráta digitálních informací, které nebudou nikdy podchyceny a archivovány. Technologie pro logickou dlouhodobou ochranu digitálních dat se teprve vyvíjejí a zdaleka nepředstavují hotové řešení. Paměťové instituce na celém světě mají problém vypořádat se s tímto novým typem dokumentů a držet tempo s technologickým pokrokem. Přerod knihoven, archivů a muzeí v digitální instituce (tzv. *going digital*) vázne.

Tento problém se ovšem netýká pouze paměťových institucí, možná ještě palčivější je v oblasti zdravotnictví (uchování digitálních dat o pacientech, vývoji léků, testech), průmyslové výrobě (opět uchování dokumentace, testování) a také v oblasti *e-governmentu*, který je na vzestupu ve vyspělém světě a také v ČR. Ve vyspělých zemích se také řeší uchování tzv. vědeckých dat, která vznikají jako výstupy různých měření, testů, výzkumů. Mnohokrát jde o data, která nelze získat znovu a mají tak velikou cenu (měření klimatu, data ze satelitů apod.). U vědeckých dat se navíc většinou jedná o obrovské objemy dat (viz hadronový urychlovač v CERN). Věda jako celek zažívá posun od experimentů prováděných v laboratoři nebo v reálném prostředí (*in vitro*) k experimentům založeným pouze na datech a jejich vyhodnocování (*in silico*). Tento trend, který se v USA nazývá někdy také *e-science* nebo *cyberscholarship*, produkuje kvanta dat a závisí na nich a na jejich uložení. Data vygenerovaná konkrétním experimentem nemohou často být vytvořena znovu (experiment nelze opakovat), je potřeba je tedy uchovat. Posun k ochraně vědeckých dat byl mj. zvláště markantní na konferenci iPRES 2011 (*International Conference on Preservation of Digital Objects*), která se konala v listopadu 2011 v Singapuru. Úvodní příspěvek, který měl Seamus Ross [ROSS, 2011], tento směr jasně naznačil. Dlouhodobou ochranou vědeckých dat se dlouhodobě mezi lety 2008-2010 zabýval např. evropský projekt *PARSE.Insight*[32] (*Permanent Access to the Records of Science in Europe*).

Pohledy na dlouhodobou ochranu v různých odvětvích i zdánlivě podobných institucí se často liší. V knihovnách se logická dlouhodobá ochrana řeší pro digitalizované dokumenty, nověji také pro digital-born dokumenty. Hlavním cílem je archivace tzv. archivní kopie (*master copy*) a hlavní

[32] Šlo o projekt ze sedmého projektového rámce (FP7) – viz http://www.parse-insight.eu/.

důraz je na dokumentaci vzniku digitálního objektu a uložení odpovídajících metadat, která zajistí zpřístupnění a pochopení dokumentů i v budoucnu. Naproti tomu např. v archivech je daleko větší důraz na informace o původci, prokázání původce (elektronické podpisy), dokumentaci procesu výběru a také integritu a autenticitu. Integrita a autenticita je důležitá v dlouhodobé ochraně vždy, ovšem v archivech je naprostou podmínkou. Velmi často, na rozdíl od knihoven, jsou ukládány jednotliviny, tj. jediné kopie konkrétního dokumentu, který často má právní a jinou působnost.

Digitální dokumenty se pravidelně a stále více objevují od konce 90. let 20. století také v archivnictví. Velká Británie, Německo a další vyspělé země začaly problematiku řešit a dospěly k závěru, že je nutné vytvořit podmínky pro dlouhodobé uložení digitálních archiválií. Že jde opravdu o nutnost je jasné z nárůstu počtu digitálních dokumentů ve státní správě, nařízení používání elektronických systémů spisových služeb (novela archivního zákona z roku 2009), využívání datových schránek také od roku 2009 apod. Pokud státní správa zažívá podobný typ změny, je nutno provést změnu také v koncepci zpracování a uložení nových typů dokumentů tak, aby digitální byl celý životní cyklus dokumentů a plně tak odpovídal životnímu cyklu klasických papírových archiválií. Nejde přitom pouze o státní správu, proměna se týká běžného života všech občanů, kdy se reálně počítá s ještě větší elektronizací administrativních úkonů než doposud (např. elektronické daňové přiznání, různé registry a jejich propojení, elektronické volby, elektronické zadávání veřejných zakázek, datové schránky atp.). Řešením nastíněných problémů se koncepčně zabývá již od roku 2004 Národní archiv a Oddělení archivní správy při ministerstvu vnitra ČR, které prosazuje novely zákonů akcentující proměny a manipulaci s digitálními dokumenty (např. vládní usnesení č. 11/2004 o dlouhodobém uchovávání a zpřístupňování dokumentů v digitální podobě; vyhláška č. 191/2009 Sb. o spisové službě aj.). Do Národního archivu v Praze od konce 90. let minulého století přicházejí dokumenty, které jsou mimo obecný rámec českého archivnictví a archiváři nejsou schopni je zpracovat, bezpečně uložit ani využít pro badatele. Jde např. o výstupy z různých informačních systémů v proprietárních formátech, nebo na historických médiích. Příkladem může být kniha došlé pošty, kterou Národní archiv ČR převzal v roce 1999 od Úřadu vlády ČR, který vedl tuto knihu od roku 1998 pouze elektronicky. Dokumenty byly předány na 78 osmipalcových disketách, na kterých byly naprosto neznámé formáty dat. Tyto dokumenty se již velmi pravděpodobně nepodaří obnovit [CROSSCZECH, [2010], s. 24]. Národní archiv a jeho pracovníci byli jedněmi z prvních, kteří v českém prostředí pojmenovali problém dlouhodobé ochrany digitálních dokumentů, a to v souvislosti s novým archivním zákonem č. 499/2004 a jeho důsledky. Již v roce 2005 prezentovali na konferenci *Archivy, knihovny a muzea v digitálním světě* příspěvek k tomuto tématu, kde se také mluví o referenčním rámci OAIS – viz kapitola 3.3. V roce 2005 také vzniklo specializované oddělení při Národním archivu [KUNT, 2006, s. 33].

3.2.4 Vývoj pohledu a přístupu k problematice dlouhodobé ochrany digitálních dat

V paměťových institucích se ve 20. století objevovaly problémy s různými typy nosičů (pásky, disky) a nemožností je přehrát nebo s jejich poškozením. Problematika dlouhodobé ochrany digitálních dat, jak ji chápeme dnes, se jako problém v paměťových institucích objevila až později, několik let poté, co se začalo ve velkém s digitalizací. Od počátku se mezi vším nadšením z digitalizace objevovaly pochyby o dlouhodobém uložení a budoucí použitelnosti dat vzniklých

digitalizací. Rodí se pojem „*digital preservation*“. Na začátku 90. let 20. století začínají vycházet první práce, které řeší problémy spojené s novými technologiemi a uchováním dat, většinou, ovšem ne výhradně, z pohledu HW a úložných médií – např. [LESK, 1992]. Již v roce 1990 zpráva amerického kongresu uváděla několik příkladů, kde velmi významné dokumenty státní správy v digitální podobě byly již ztraceny díky nepřipravenosti institucí na problematiku dlouhodobé ochrany [ROTHENBERG, 1999, s. 4].
V oblasti paměťových institucí si nebezpečí mezi prvními uvědomila Australská národní knihovna, která začala ihned konat. V roce 1993 začala budovat pracovní skupinu *Preserving Access to Digital Information*, která začala vytvářet a posléze aktualizovat i velmi významnou webovou stránku, pod zkratkou PADI[33]. Web existoval mezi lety 1997-2010, kdy byl zrušen pro nedostatek financí. Za celou dobu své existence byl velmi významným zdrojem informací o dlouhodobé ochraně digitálních dat. Informace byly pravidelně doplňovány a byly rozděleny podle jednotlivých témat.[34] Téma ochrany digitálních dat se poté občas objevovalo na různých konferencích, věnovaných převážně digitalizaci. Např. konference *Preservation and Digitisation – principles, practice and policies*, která se konala v roce 1996 ve Velké Británii, byla věnována digitalizaci, ale v úvodu ke sborníku se mluví také o „*digital preservation*“. Dva příspěvky konference tomuto problému byly přímo věnovány. Jednalo se o *Long-term Preservation of Digital Materials* (J.W.C. van Bogart) a *Digital Archiving* (D. J. Waters) – viz [BRITISH LIBRARY. NATIONAL PRESERVATION OFFICE, 1997]. Pojem „*digital preservation*“ je v úvodu vysvětlován ve dvou smyslech, jako ochrana fyzických předloh pomocí digitalizace a také ve smyslu, který dnes naprosto převládá, tedy jako dlouhodobá ochrana digitálních dat [BRITISH LIBRARY. NATIONAL PRESERVATION OFFICE, 1997, s. 1], i když v tehdejším pojetí šlo stále spíše o ochranu bitstreamu, tedy fyzických nosičů jako jsou pásky a disky, tedy ne logickou dlouhodobou ochranu.

Velmi důležitou, z mého pohledu průlomovou, v artikulaci problémů a formování pochopení dlouhodobé ochrany byla americká zpráva *Preserving Digital Information*, která vznikala z popudu organizací *The Research Library Group* a *The Commission on Preservation and Access* od roku 1994 a byla vydána v roce 1996 [GARRETT a WATERS, 1996]. Jako jeden z prvních dokumentů definovala dlouhodobou ochranu digitálních dat v dnešním smyslu, popsala problematiku zastarávání SW i HW, možnosti řešení (migrace), otázky integrity digitálních dokumentů, jejich provenience a zachycení kontextu atp. Zabývala se také podrobně digitálními repozitáři. Závěr této zprávy byl šokující, autoři konstatovali, že v době vydání zprávy neexistovala žádná vhodná metoda na dlouhodobou ochranu a společnost a její instituce nejsou na tuto problematiku připraveny. Konstatovali nepřipravenost na legislativním poli i v otázce odpovědnosti jednotlivých institucí.

Ve Velké Británii odstartoval v roce 1998 projekt CEDARS (*CURL Exemplars in Digital ARchiveS*), jehož cílem bylo propagovat možná řešení dlouhodobé ochrany digitálních dat, vytvářet a zpřístupňovat různé strategie a návody. Projekt řešil tři roviny dlouhodobé ochrany, jak byla chápána v roce 1998: ochranu média; ochranu integrity digitálního objektu a řešení zastarávání SW a HW [RUSSELL, 1998]. V roce 1999 se konal workshop na téma dlouhodobé ochrany digitálních dat na univerzitě ve Warwicku, Velká Británie. Workshop byl jen pro zvané účastníky a

[33] http://www.nla.gov.au/padi/index.html
[34] Nyní se zdá, že s pomocí sdružení ICADS (IFLA-CDNL *Alliance for Digital Strategies*) je web PADI opět aktivní a je aktualizován.

jeho cílem bylo vytvořit agendu a doporučení pro následujících pět let. Tento workshop v konečném důsledku vedl k založení organizace *Digital Preservation Coalition* (DPC)[35], která je velmi aktivní na poli dlouhodobé ochrany dodnes. Cílem DPC bylo od počátku řešit dlouhodobou ochranu digitálních dat na národní úrovni a podporovat mezinárodní spolupráci [WALKER, 2006, s. 542].
V té době ještě nebyl k dispozici žádný nástroj, který by dlouhodobé ochraně významně pomohl, natož komplexní nástroj typu dnešních LTP systémů. V roce 1999 se objevil první draft referenčního rámce OAIS (*Open Archival Information System*), který dal základ dalšímu vývoji v následujících letech – viz kapitola 3.3.

Až do konce tisíciletí šlo ale o období, které se logicky zaměřovalo na popularizaci problematiky dlouhodobé ochrany, spíše než na vývoj nástrojů a další aktivity. Během tohoto období si museli odborníci z jednotlivých institucí, často ve spolupráci, uvědomit, co vše je nutné provést k zachování digitálních sbírek a jejich použitelnosti v budoucnu. Pak o tom museli přesvědčit odpovědné osoby v managementu svých institucí a až poté další instituce ve své zemi i jinde. Problematika dlouhodobé ochrany digitálních dokumentů se začala více řešit na přelomu tisíciletí, kdy knihovny i archivy musely začít reagovat na skutečnost, že do nich začaly přicházet digitální objekty, které neměly fyzickou předlohu, tzv. digital-born dokumenty. Můžeme říci, že to byl i jeden z podnětů pro urychlení prací na vývoji odpovídajících systémů na dlouhodobou ochranu (LTP systémů). Úspěšný byl britský projekt CAMiLEON (1999-2003), v rámci kterého vzniklo několik strategií ochrany, byla vyřešena otázka opětovného zpřístupnění tzv. *Domesday Book* a hlavně se problematika ochrany digitálních dat stala diskutovanou ve Velké Británii i jinde. A to díky publicitě, kterou zajistil dnes asi jeden z nejznámějších příkladů aplikace opatření dlouhodobé ochrany digitálních dat, *BBC Domesday Book project*. V roce 1986 vznikl *Domesday Book* projekt, který měl za cíl zaznamenat během dvou let všední život ve Velké Británii, podobě jako jeho středověká předloha z roku 1096. Do projektu svými texty a fotografiemi přispělo více než milion lidí. Všechna data byla uložena na tehdejší technickou novinku, na speciální disky (Laser Discs; LV-ROM) a byl vytvořen speciální SW na jejich prohlížení. Technologie laser disků se ovšem v následujících letech neujala a bylo jasné, že s dobou budou data uložená na neobvyklých nosičích zobrazitelná pouze s proprietárním SW více a více ohrožena. V roce 1999 se začalo s řešením v projektu CAMiLEON. Byl vytvořen emulační i migrační nástroj k opětovnému oživení dat – více viz [ABBOTT, 2003]. Celá záchranná akce trvala dva roky.

V následujícím desetiletí vývoj v oblasti dlouhodobé ochrany digitálních dat oproti předchozímu období zrychlil, ale stále převládal, zvláště mezi lety 2000-2005, pesimistický pohled na rychlost postupu směrem k řešení a identifikaci odborné komunity s problémem. Seamus Ross v roce 2004 vyjadřuje potěšení nad pokrokem v rozšíření povědomí o problematice, ale zároveň zklamání nad tím, že stále velmi málo institucí se tématem zabývá ve svých projektech digitalizace. Neuvědomují si, že digitalizace je jen počátek řešení [ROSS, 2004 s. 92 a 95]. V roce 2004 vydali Brian Lavoie a Lorcan Dempsey článek *Thirteen ways of looking at digital preservation*, ve kterém shrnuli celou problematiku dlouhodobé ochrany ze svého pohledu [LAVOIE a DEMPSEY, 2004]. Dlouhodobou ochranu považovali na poli digitálních informací za nejméně vyvinutou oblast, i díky tomu, že se řešila pouze z pohledu hledání technického řešení, tedy izolovaně od ostatních

[35] http://www.dpconline.org/

problémů. Upozorňují, že směr řešení se pomalu otáčí k problematice *digital curation*, tedy k širším souvislostem zásad tvorby a správy digitálních dat během celého životního cyklu, což se později ukázalo jako jediná možná cesta a chápání problému dlouhodobé ochrany, které vede k výsledkům. Autoři konstatují, že často zájem vytvořit více digitálních dokumentů převáží nad zájmem ty stávající ochraňovat. Tato skutečnost byla problémem, ovšem v nových projektech EU z posledních let je znát jasný trend podporovat dlouhodobou ochranu i na úkor digitalizace a z projektů digitalizace podporovat pouze takové, které mají ochranu vyřešenu, nebo je jejich součástí. Zároveň ve svém článku apelují na kooperaci na tomto poli, což se v současnosti děje. Žádná instituce si nedovolí řešit dlouhodobou ochranu izolovaně od ostatních, znalosti i zdroje jsou často sdíleny.

Po roce 2005 se problematika dlouhodobé ochrany zcela vyčleňuje z oblasti ochrany tradičních (tedy fyzických) knihovních fondů a stává se z ní samostatná disciplína, která má svoje metody, terminologii, publikace, konference a je potřeba mít odlišné znalosti k jejímu provádění, než bylo nutné k ochraně fyzických sbírek. V období po roce 2005 proběhlo a stále probíhá mnoho dotazníkových akcí, jejichž iniciátory jsou většinou různé projekty EU, které mají za úkol zjistit stav řešení a pohledu na problematiku dlouhodobé ochrany digitálních dat v Evropě a často i ve světě. Uvést můžeme dotazníkovou akci projektu DPE z let 2006 a 2007 [STOKLASOVÁ a HUTAŘ, 2007, s. 87-88] a také výstup projektu reUSE z roku 2005 [KRIMBACHER, NEUHAUSER a VOGL, 2005], který měl podobné otázky i výsledky jako v projektu DPE. Výsledky reUSE ukázaly mj., že 76% respondentů považovalo dlouhodobou ochranu digitálních dat za velmi důležitou, přičemž nejmenší procentuální podpora tohoto názoru byla v nových členských státech EU (65%), nejvyšší ve Velké Británii a Irsku (89%). Velmi znepokojujícím faktem bylo, že pouze 20% knihoven mělo v roce 2005 strategii dlouhodobé ochrany, a z nich pouze 9% ji publikovalo online.
Začaly vznikat první specifikace metadat určených k podpoře procesů dlouhodobé ochrany digitálních dat. Prvním významným schématem byl návrh *Preservation Metadata: Metadata Implementation Schema* z novozélandské národní knihovny, jehož první verze vznikla koncem roku 2002, finální verze v červenci 2003. Na tato metadata navázalo schéma LMER (*Long Term Preservation Metadata for Electronic Resources*) německé národní knihovny, které bylo publikováno ve verzi 1.0 v roce 2004 a ve finální verzi 1.2 pak v roce 2005. V roce 2005 taky vznikla první verze datového slovníku a schématu PREMIS (*Preservation Metadata: Implementation Strategies*) [OCLC/RLG PREMIS WORKING GROUP, 2005], který od té doby oblasti ochranných metadat dominuje a je dnes nezbytnou součástí všech procesů digitalizace a dat určených pro dlouhodobé uložení a ochranu. Cílem bylo poskytnout základní sadu ochranných metadat.
Objevily se první projekty a první verze nástrojů, jako např. nástroje na extrakci metadat (2004 - JHOVE, New Zealand Metadata Extractor), na normalizaci dat (XENA v NK Austrálie), později registry formátů jako je např. PRONOM[36] a UDFR[37] (*Unified Digital Formats Registry*), které staví právě na PRONOMu a svém předchůdci GDFR[38] (*Global Digital Formats Registry*). Ve Velké Británii vzniklo *Digital Curation Centre* (DCC[39]), které hrálo a stále hraje nezastupitelnou roli v popularizaci dlouhodobé ochrany, vytváření návodů, odborných článků a školení.

[36] http://www.nationalarchives.gov.uk/pronom/
[37] http://www.udfr.org/
[38] http://www.gdfr.info/
[39] http://www.dcc.ac.uk/

Začala se také reálně řešit certifikace repozitářů, první pokus o specifikaci kritérií nutných k certifikaci vydala v roce 2002 *Research Libraries Group* [RESEARCH LIBRARIES GROUP, 2002], což byl první krok ke vzniku certifikační metodiky *Trustworthy Repositories Audit & Certification* známé dnes pod zkratkou TRAC [OCLC, CRL, 2007] – více viz kapitola 6.4.2.

I přes všechny projekty a postup v oblasti dlouhodobé ochrany digitálních dat, konstatoval jeden ze základních a výchozích dokumentů projektu DPE, tzv. roadmapa, že ani po 20 letech výzkumů se znalosti komunity, pozice a východiska v oblasti ochrany digitálních objektů výrazně oproti době před 20 lety nezlepšily. Na základě analýzy probíhajících projektů a výzkumů zpráva konstatovala, že komunita sdružená okolo „*digital preservation*" se problému ochrany digitálních objektů stále nezhostila odpovídajícím a dostatečným způsobem [DIGITAL PRESERVATION EUROPE, 2007, s. 8]. Důvodem takového konstatování byla skutečnost, že široce aplikovatelná řešení k problematice ochrany digitálních dokumentů byla spíše výjimkou než pravidlem. Od roku 2007 se ale mnohé změnilo.

Reálné systémy, které lze považovat za funkční LTP z dnešního pohledu, začaly vznikat až v druhé polovině první dekády nového tisíciletí (KRONOS[40] v národní knihovně Nového Zélandu, Safety Deposit Box ve firmě Tessella). Vyvíjena jsou také open source řešení, která se objevují po roce 2009 (např. Archivematica, E-prints, Hoppla, Mopseus, RODA), jež jsou dnes k dispozici nejen knihovnám, muzeím a archivům [FOJTŮ, HUTAŘ a MELICHAR, 2011, s. 73-74]. Nejaktivnější ve výzkumu a implementaci logické dlouhodobé ochrany digitálních dat jsou tradičně knihovny a archivy ve Velké Británii, Německu, Nizozemí, Austrálii a na Novém Zélandu, v severských zemích, USA a Kanadě. Přidávají se ostatní evropské a světové knihovny (Singapur, Francie, Estonsko, Polsko, Slovensko).

3.2.4.1 Projektová a jiná podpora dlouhodobé ochrany digitálních dat

Problémy s dlouhodobou ochranou digitálních dat se začaly objevovat v souvislosti s digitalizací a první koncepty, které ji alespoň částečně braly v potaz, se týkaly primárně digitalizace (*Parmská charta, Lundský plán* apod.). Projekty věnující se dlouhodobé ochraně plynule navazují na projekty digitalizace. Jednou z prvních snah EU na poli dlouhodobé ochrany digitálních dat bylo zvýšení povědomí o problému, které by tak nastartovalo vlastní výzkum.

Zástupcem prvních projektů byl evropský NEDLIB (*Networked Europian Deposit Library*) z programu *Telematics for the Libraries*, který měl za cíl navrhnout systém na dlouhodobou ochranu dle specifikace OAIS. Projekt běžel v letech 1998-2000. Přínosem bylo rozsáhlé testování emulace jako jedné z metod dlouhodobé ochrany digitálních dokumentů. V roce 2001 odstartoval důležitý projekt ERPANET (*Electronic Resource Preservation and Access Network*). Cílem projektu bylo vytvořit evropské konsorcium k propagaci informací, best practices a dovedností pro oblast dlouhodobé ochrany kulturních a vědeckých digitálních objektů. Za dobu trvání projektu (2001-2004) se podařilo shromáždit velké množství instruktážních materiálů, návodů, studií, pořádaly se workshopy, konference apod. Web ERPANET[41] dodnes zůstává významným informačním zdrojem, na který plynule navázal web projektu DPE (*DigitalPreservationEurope*[42]), který měl velmi podobné cíle jako ERPANET. Tedy usnadnit sdílení výsledků výzkumů, které existují napříč vědeckými, akademickými, kulturními, veřejnými a oborovými sektory v Evropě a podporovat

[40] Později přejmenován na LTP systém Rosetta.
[41] http://www.erpanet.org/
[42] http://www.digitalpreservationeurope.eu/

spolupráci a součinnost mezi četnými již existujícími národními iniciativami [HUTAŘ a NERGLOVÁ, 2007].

Začátek vlastních výzkumů a vědecké práce spočíval v definování konkrétních problémů, terminologie. Vznikly první nástroje (např. v projektech NEDLIB a DELOS), aktivity zaměřené na metadata, výběr dat, identifikaci formátů apod. Výzkum byl zaměřen převážně na „kancelářské" dokumenty a obrazová data v institucích. Až později se řešila problematika ochrany komplexních a netradičních dokumentů. V roce 2004 volal Maurizio Lunghi, tehdy koordinátor tzv. *Florentské pracovní skupiny*, po tom, aby se dlouhodobá ochrana digitálních dat řešila koordinovaně v rámci EU a v souvislosti s projekty digitalizace, aby se vyčlenilo financování pouze na tuto oblast a aby se přešlo od výzkumu k činům [LUNGHI, 2004, s. 85-86]. Florentská agenda vznikla za italského předsednictví EU a byla vytvořena skupinou odborníků za podpory projektů MINERVA a PRESTOSPACE. Obsahovala tři akční body: 1) zvýšit u odpovědných pracovníků světových knihoven, archivů a muzeí povědomí o rizicích a hrozbách plynoucích z ukládání digitálních dokumentů a zvýšit úroveň spolupráce na řešení; 2) vyhledání a popis iniciativ a dostupných nástrojů na dlouhodobou ochranu digitálních objektů; tvorba tzv. *good practice* a prezentací i školení; 3) právní implikace a problémy s problematikou spojené – více viz [The Firenze Agenda, 2003, s. 29].
Apely skupiny pro zařazení problematiky dlouhodobé ochrany digitálních dat do financování EU, se promítly do iniciativy i2010: Evropská informační společnost pro růst a zaměstnanost (*i2010: A European Information Society for Growth and Employment*), respektive *i2010: Digitální knihovny* (*i2010: Digital Libraries*) z roku 2005. Jako hlavní cíle „Iniciativy Digitální knihovny" byly stanoveny online dostupnost zdrojů, nutnost digitalizace tradičních sbírek a jejich ochrana a dlouhodobé uchovávání, což se promítlo do evropských projektů od rámcového programu 6 (FP6).
Dlouhodobé ochrany se týkaly mj. tyto projekty: FP5 (1998-2001; projekty ERPANET, DigiCULT, PRESTO, ECHO, MINERVA, FP6 (2002-2006; PRESTOSPACE, DELOS, BRICKS, CASPAR, DPE, PLANETS, PRESTOSPACE) a FP7 (2007-2013; PROTAGE, DL.ORG, LIWA, SHAMAN, KEEP, PRESTOPRIME, SCAPE, ARCOMEM, TIMBUS). Projekty z FP6, které jsou již nyní ukončeny, poskytly mnohá technická řešení a nástroje. Zvláště úspěšné ve vývoji byly projekty PLANETS[43] a CASPAR[44], nebo také KEEP[45].
Projekt PLANETS[46] (*Preservation and Long-term Access through NETworked Services*) byl více prakticky zaměřený než DPE.[47] Cílem bylo vytvoření nástrojů pro plánování dlouhodobé ochrany a testovací prostředí pro migrace a jiná opatření prováděná na digitálních objektech. Projektu se navíc účastnily firmy, které nabízejí komerční řešení pro LTP systém (Tessella, UK). Výstupy projektu jsou pro odbornou komunitu velmi podstatné. Jedná se o sadu PLANETS nástrojů, která obsahuje PLATO, PLANETS testbed, GRATE emulátor (*Global Remote Access to Emulation Services*), sada SIARD (vše viz kapitola 3.2.5). Všechny nástroje jsou volně dostupné, některé jako online aplikace a jsou volně k využití. Projekt PLANETS byl velmi úspěšný a vytvořil okolo sebe komunitu odborníků, kteří chtěli spolupracovat a rozvíjet nástroje i po ukončení projektu. Vznikla

[43] http://www.planets-project.eu/
[44] http://www.casparpreserves.eu/
[45] http://www.keep-project.eu/ezpub2/index.php
[46] http://www.planets-project.eu/
[47] Personálně i institucionálně se oba projekty prolínaly. Díky tomu měl tým NK ČR, který se účastnil DPE, velmi dobré vztahy s projektem PLANETS i s jeho řešiteli.

proto *Open Planets Foundation*[48] (OPF) v rámci které jde vývoj dál. Vývoj započatý v projektech PLANETS a CASPAR pokračuje v projektech SCAPE a SHAMAN (oba FP7). SCAPE má zlepšit stav výzkumu dlouhodobé ochrany vývojem infrastruktury a nástrojů pro škálovatelnou ochranu, poskytnutím rámce pro automatická workflow. Všechny vyvinuté komponenty budou zapojeny do systému na dlouhodobou ochranu. Projekt SHAMAN vyvíjí novou generaci nástrojů pro analýzu, ingest, správu a zpřístupnění digitálních objektů.

Jak vyplývá z přehledu EU výzkumu na poli dlouhodobé ochrany digitálních dat z roku 2011 [STRODL, PETROV a RAUBER, 2011], bylo na tuto problematiku vynaloženo skoro 100 milionů Euro a to do více než patnácti projektů v obou rámcových programech FP6 (2002-2006) a FP7 (2007-2013). Je důležité si také uvědomit, že financování FP7 se oproti původním plánům více než ztrojnásobilo, což ukazuje na to, jaký důraz EU problému věnuje. Díky výstupům z těchto projektů máme k dispozici dnešní nástroje, znalosti a také mnoho příkladů použití v praxi. Tyto projekty měly přesah i na mezinárodní snahy (TRAC certifikace, referenční rámec OAIS, PREMIS) [STRODL¨, PETROV a RAUBER, 2011].

V současnosti se aktivity na poli dlouhodobé ochrany dají shrnout pod hesla základní a aplikovaný výzkum. Základní výzkum jde za hranici jednoduchých digitálních objektů. Důraz je na interaktivních objektech, objektech vložených apod. Příkladem může být projekt LiWA (*Living Web Archives*), kterého se účastnila i NK ČR, a který se zabývá archivací dat ze sklízení webu, která jsou velmi komplexní. Podobně projekt TIMBUS, který začal v roce 2011, se věnuje výzkumu na poli ochrany business procesů. Základní výzkum dále pokračuje také na úrovni formálních metod pro validaci objektů v projektech PLANETS a nově SCAPE [STRODL, PETROV a RAUBER, 2011, s. 3].

Aplikovaný výzkum se v národních a evropských projektech věnuje vývoji škálovatelných systémů na dlouhodobou ochranu. Komunita potřebuje nástroje, metody i celé systémy k ostrému nasazení a využívání pro různorodé a velmi rozsáhlé sbírky digitálních objektů. S tím se pojí automatizace, rozhodovací procesy, výkon nástrojů, kriteria validace. V minulosti byly nástroje a moduly pro dlouhodobou ochranu navrženy tak, že vyžadovaly lidské řízení. Nyní je snaha provádět rozhodnutí i procesy automatickou cestou. Příkladem může být projekt SCAPE a také projekt ARCOMEM, který využívá sociální weby pro automatickou tvorbu informací a podporu výběru [STRODL, PETROV a RAUBER, 2011, s. 3].

Budoucnost vývoje na poli dlouhodobé ochrany digitálních dat se může ubírat směrem ven od paměťových institucí, např. na stranu komerčních firem, *e-govermentu*, *e-health*, *e-commerce* apod. Tyto instituce a oblasti si začínají uvědomovat palčivost ochrany digitálních objektů až nyní.

Z projektů nefinancovaných z EU peněz je nutno jmenovat německý *Nestor*[49] a americký *National Digital Information Infrastructure and Preservation Program* (NDIIPP[50]), který je asi největším počinem v oblasti ochrany digitálních dat mimo EU. Dodnes fungující projekt NDIIPP začal v roce 2000, kdy si Kongresová knihovna určila za cíl vyvinout národní strategii pro ochranu digitálních informací. Americký kongres vyčlenil dotace ve výši 100 milionů dolarů, uvolňovaných po jednotlivých letech. V první fázi příprav do roku 2003 byl vytvořen plán ve spolupráci s různými institucemi a producenty dat. Od roku 2003 se začalo s projektem podle vytvořeného plánu. Od roku 2005 v rámci programu existovala speciální část věnovaná výzkumu na poli dlouhodobé ochrany. Program byl mj. odpovědí na projekty digitalizace, na které bylo relativně lehké získat

[48] http://www.openplanetsfoundation.org/
[49] http://www.langzeitarchivierung.de/eng/
[50] http://www.digitalpreservation.gov/

finanční podporu, instituce ale potřebovaly platformu, na které by mohly stavět procesy dlouhodobé ochrany. Program probíhá dodnes a to ve spolupráci s knihovnami, archivy, univerzitami v USA.

Vzhledem k tomu, že řešení dlouhodobé ochrany digitálních dat není v silách jediné instituce, vzniklo za posledních 15 let vedle projektů také několik koordinačních a podpůrných institucí. Ve Velké Británii existuje *Digital Preservatin Coalition* (DPC[51]), *Joint Information Systems Committee* (JISC[52]), *Digital Curation Centre* (DCC[53]). V Německu *Nestor* (*Network of Expertise in Long-Term Storage of Digital Resources*), v USA *National Digital Information Infrastructure and Preservation Program* (NDIIPP), v Nizozemí DANS[54] a od roku 2008 NCDD[55] (*The Netherlands Coalition of Digital Preservation*).
Existují také konference, jako např. iPRES[56], Archiving[57], JCDL[58], ECDL/TPDL[59] a odborné časopisy podporující sdílení znalostí v této oblasti (*International Journal of Digital Curation*[60], *Ariadne*[61], *D-Lib Magazine*[62] aj.). Nejvýznamnější z uvedených konferencí je iPRES (*International Conference on the Preservation of Digital Objects*), která se zabývá pouze a výlučně dlouhodobou ochranou digitálních dat. První ročník proběhl v roce 2004 v Pekingu, poslední na podzim 2011 v Singapuru. Koná se každý rok minimálně po tři dny, které jsou naplněny tutorialy, workshopy a přednáškami.

3.2.4.2 Dlouhodobá ochrana digitálních dat v NK ČR

Období počátku digitalizace ve světě i v ČR bylo jednoznačně ve znamení popisných metadat a výzkumu v oblasti digitalizace. V České republice ani v NK ČR se proto aktivně logická dlouhodobá ochrana digitálních dat dlouho neřešila, ačkoliv o této problematice určitá povědomost byla relativně brzy.[63] V různých článcích na konci 90. let 20. století je možné pozorovat náznaky, zmínky o ochranných metadatech, případně o hrozbách spojených se zastaráváním SW i HW – viz např. [KNOLL, 1997]. Podobným článkem je i *Problematika elektronických publikací* stejného autora [KNOLL, 1999], který pojednává o metadatech. Mimo popisná metadata jsou zmíněna také metadata, která mají zajistit udržení integrity digitálních objektů a také jejich kontext, tedy informací nutných pro jeho zpřístupnění v budoucnu. Technická metadata jsou nahlížena z pohledu uživatele, mají mu zajistit věrné zobrazení. Integrita je uvažována z pohledu kompletnosti, ne integrita jednotlivých digitálních objektů. Součástí článku je i kapitola Ochrana a dlouhodobé zpřístupnění, která popisuje možná řešení (HW a SW muzeum, emulace a migrace). Autor vyjádřil svůj názor, že migrace je v podstatě neproveditelná finančně i pracovně [KNOLL,

[51] http://www.dpconline.org/
[52] http://www.jisc.ac.uk/
[53] http://www.dcc.ac.uk/
[54] http://www.dans.knaw.nl/en/content/about-dans
[55] http://www.ncdd.nl/en/index.php
[56] http://www.ifs.tuwien.ac.at/dp/ipres2010/
[57] http://www.imaging.org/ist/conferences/archiving/
[58] http://www.jcdl.org/
[59] http://www.tpdl2011.org/
[60] http://www.ijdc.net/index.php/ijdc
[61] http://www.ariadne.ac.uk/
[62] http://dlib.org/
[63] Dodnes je bohužel chápání dlouhodobé digitální ochrany v mnohých institucích omezeno na ochranu bitstreamu.

1999, s. 173] a naděje vkládá do emulace. Dnes víme, že migrace je nejužívanější metodou dlouhodobé ochrany a to i díky rozvoji technologií.
Rok na to, v roce 2000, Filip Vojtášek publikoval v časopisu Ikaros článek s názvem *Dlouhodobá archivace digitálních dokumentů*, kde jako první z autorů explicitně rozebírá základní otázky logické ochrany digitálních dat pro budoucnost [VOJTÁŠEK, 2000]. Obavy o budoucnost digitálních dokumentů se objevovaly i nadále. V roce 2002 Jiří Polišenský na semináři CASLIN ve svém příspěvku mluví o tom, že: „*Digitální dokumenty je na rozdíl od mikrofilmu, mnohem obtížnější uchovat ve zpřístupnitelné podobě v dlouhodobé časové perspektivě, což je dáno vývojem v oblasti výpočetní techniky a … inovačními cykly přinášejícími zásadní změny v technické podpoře platforem, formátů…*“ [POLIŠENSKÝ, 2002, s. 57] Situace ale ještě nebyla taková, aby tyto obavy vyústily do konkrétních opatření. Lze to přičíst obecné úrovni znalostí o logické dlouhodobé ochraně v té době, stav poznání byl omezený, dnes běžné technologie a nástroje neexistovaly, např. služby, které pomáhají s identifikací formátů, jejich validací, obohacením metadat apod. (JHOVE, PRONOM/DROID aj.). Tyto nástroje se objevují až po roce 2005. V době specifikace standardů metadat pro aplikace Kramerius a Manuscriptorium také neexistovaly systémy na dlouhodobou ochranu digitálních dat, které dnes existují v komerčních i open source verzích. V nizozemské národní knihovně byl sice již od roku 1999 vyvíjen a v roce 2003 plně funkční LTP systém DIAS. Ovšem povědomí o jeho skutečné funkcionalitě a implikacích pro NK ČR bylo velmi malé.
Až po roce 2005 se stávala problematika logické dlouhodobé ochrany ve světě stále více aktuální. Docházelo k reálnému nasazení a rozšíření tzv. ochranných metadat (PREMIS), instituce zároveň začaly dělat opatření ve svých procesech, které měly logické ochraně dat napomoci. Nizozemská národní knihovna měla od roku 2003 funkční systém e-Depot, Národní knihovna Nového Zélandu vytvářela specifikaci systému a vlastní systém pod názvem Kronos (dnešní Rosetta). Bylo jasné, že novou problematiku je nutno pojmout v rámci NK ČR pojmout obšírněji a začít se jí zabývat hlouběji. Vždyť NK ČR byla jednou z mála světových knihoven, která ukládala několik milionů naskenovaných stran z předloh ze svých bohatých sbírek. Nutnost ochrany zdigitalizovaných dokumentů a jejich použitelnosti pro budoucnost byla stále jasnější. V roce 2005 tedy vznikl v NK ČR *Referát pro digitální knihovnu NFS* s jedním úvazkem podřízeným přímo ředitelce Novodobých fondů a sbírek NK ČR. Z referátu během let vznikl *Odbor digitální ochrany* s deseti celými úvazky.

Základní formulaci problémů s dlouhodobou ochranou digitálních dat a snahu o řešení lze najít ve výzkumném záměru VaV *Vytvoření virtuálního badatelského prostředí pro zpřístupnění a ochranu digitálních dokumentů*, který řešila NK ČR v letech 2004-2010, a který má dlouhodobou ochranu digitálních dat přímo v názvu. O trochu podrobněji je dlouhodobá ochrana digitálních dat a návrhy jejího řešení pojednány v rozpracované části Koncepce rozvoje knihoven ČR na léta 2004-2010 v podobě *Koncepce trvalého uchování knihovních sbírek tradičních a elektronických dokumentů v knihovnách ČR do roku 2010* [Koncepce trvalého uchování…, 2005]. Koncepce trvalého uchování jasně navazovala na iniciativu *i2010: Digitální knihovny* a obsahovala plán rozvoje v oblasti uchování tradičních i digitálních fondů včetně SWOT analýz a finančního výhledu. Koncepce velmi jasně říká, že zatímco snahy o uchování se stále soustředí na tradiční fyzické dokumenty, v podobě digitálních objektů nám uniká velká část kulturního dědictví, které není nikde jinde fixováno a je tímto ztraceno. Koncepce ovšem nejde dál, než ke konstatování, že i digitální objekty je nutné ochránit (rozuměj získávat a ukládat). K tomuto účelu bude nutné nakoupit a zprovoznit robustní systém. Centrální datové úložiště bylo později sice vybudováno a dále omezeně rozvíjeno, ale vždy

z ad-hoc přidělených peněz z Ministerstva kultury ČR, když situace uložení dat v NK ČR začala být kritická. Koncepce nezmiňuje problémy spojené s logickou dlouhodobou ochranou digitálních dokumentů.
V rámci VaV *Vytvoření virtuálního badatelského prostředí pro zpřístupnění a ochranu digitálních dokumentů* vznikla v roce 2005 analýza, která formulovala zásady dlouhodobé ochrany digitálních dokumentů i rizikové faktory [KNOLL, POLIŠENSKÝ a UHLÍŘ, 2005, s. 11]. Zásady jsou bohužel aplikovatelné na ochranu bitstreamu a neberou v potaz logickou ochranu digitálních dat. I tak lze analýzu považovat za jeden z dalších kroků na cestě k logické ochraně digitálních dat v NK ČR, protože odpovídající úroveň uložení, zálohování, udržení integrity, vytvoření migračních politik archivních dat (bitstreamu) je vždy předpokladem logické ochrany těchto dat. Ve stejném VaV byla popsána i rizika vyplývající ze zastarávání SW a HW a možná řešení pro NK ČR. Uvedena byla migrace, emulace a technologické muzeum [KNOLL, POLIŠENSKÝ a UHLÍŘ, 2005, s. 9]. Dnes je již jasné, že navrhované řešení „technologické muzeum" není řešením logické dlouhodobé ochrany digitálních dat – viz kapitola 3.2.3.

Prvním krokem směřujícím k využívání systému na logickou dlouhodobou archivaci digitálních dat bylo zakomponování plánu na pořízení takového systému do *Koncepce trvalého uchování knihovních sbírek tradičních a elektronických dokumentů v knihovnách ČR do roku 2010*, kterou schválila česká vláda. Této skutečnosti využili řešitelé VaV *Vytvoření virtuálního badatelského prostředí pro zpřístupnění a ochranu digitálních dokumentů* a v roce 2006 navrhli oblast dlouhodobé ochrany v tomto výzkumném záměru podstatně omezit. Návrh změny byl schválen v oponentním řízení [KNOLL, POLIŠENSKÝ a UHLÍŘ, 2007, s. 3] a původně plánovaná problematika dlouhodobé ochrany z VaV v následujících letech takřka vymizela, resp. omezila se jen na omezenou migraci obsahu CD/DVD disků z povinného výtisku na Centrální datové úložiště NK ČR. Řešitelé VaV počítali s tím, že HW a LTP systém se nakoupí z jiných finančních zdrojů (koncepce aj.). Ušetřené finance a pracovní nasazení byly vloženy do projektu *Manuscriptorium* a jeho dalšího rozvoje jako virtuálního badatelského prostředí [KNOLL, POLIŠENSKÝ a UHLÍŘ, 2006, s. 2 a 19]. Z jiných zdrojů se podařilo v roce 2006 nakoupit datové úložiště (více viz kapitola 6.3) financované z grantu Ministerstva informatiky ČR a zajistit tak odpovídající uložení dat, bohužel ne logickou dlouhodobou ochranu digitálních dat. NK ČR tak sice získala HW vrstvu jako základní předpoklad pro logickou ochranu dat, ale stále neměla odpovídající SW na správu a ochranu dat (LTP systém). Již v roce 2005 padlo v rámci uvedeného VaV rozhodnutí, že archivační systém se pro svou náročnost nebude vyvíjet a nakoupí se hotové řešení, které je součástí koncepce. Paradoxem celého tohoto opatření v rámci VaV je, že k naplnění výše zmíněné koncepce z roku 2005 nikdy nedošlo. I přesto, že koncepci schválila Vláda ČR, nenašly se dostatečné finanční prostředky na její naplnění. NK ČR tak odpovídající LTP systém nemá dodnes.

Přelomovým rokem pro NK ČR v oblasti logické dlouhodobé ochrany dat byl bezpochyby rok 2006. V tomto roce se NK ČR stala partnerem evropského projektu DPE (*DigitalPreservationEurope*[64]). Projekt trval od dubna 2006 do dubna 2009 a měl za cíl zvýšit povědomí o problematice logické dlouhodobé ochrany digitálních dat v evropském prostoru, identifikovat akutní problémy k řešení, zajistit spolupráci různých (nejen knihovnických) komunit, vytvořit celoevropský školící program o problematice ochrany digitálních dat atd. – viz [DIGITAL PRESERVATION EUROPE, 2006]. Cílem

[64] http://www.digitalpreservationeurope.eu/

účasti NK ČR bylo získat zkušenosti a finance na úvazky nového odboru a také aplikovat jeden z cílů projektu, zvýšení povědomosti o problému dlouhodobé ochrany, v ČR. Zprvu byli řešitelé za NK ČR Mgr. Adolf Knoll a PhDr. Bohdana Stoklasová, později od roku 2007 Mgr. Jan Hutař a PhDr. Bohdana Stoklasová (ředitelka úseku *Novodobých fondů a sbírek* NK ČR). Projekt DPE byl významný díky několika skutečnostem. Pracovníci NK ČR se dostali do komunity odborníků, kteří se logickou ochranou digitálních dat reálně a na světové úrovni zabývali ve svých institucích; tj. NK ČR měla možnost účastnit se řešení aktuálních problémů na mezinárodní scéně. Spolupráce s DPE také vyústila ve spolupráci se dvěma dalšími komunitami okolo evropských projektů CASPAR[65] a PLANETS[66]. V říjnu 2008 proběhl v NK ČR týdenní workshop WePreserve/DRAMBORA, ve spolupráci s odborníky z výše jmenovaných projektů. Zúčastnili se zástupci mnoha českých i zahraničních institucí, mnozí se o dlouhodobé ochraně digitálních dat dozvěděli vůbec poprvé[67]. Workshop měl velký ohlas v komunitě českých paměťových institucí. Během projektu DPE v NK ČR také bylo publikováno množství studií v českém jazyce[68] a také publikace o plánování digitálních repozitářů PLATTER, jejímž spoluautorem byl autor této disertační práce [ROSENTHAL, BLEKINGE-RASMUSSEN a HUTAŘ, 2009]. PLATTER byl prvním českým uceleným dokumentem, který vysvětloval logickou dlouhodobou ochranu digitálních dat. V rámci projektu DPE také probíhal vývoj metodiky a nástroje na audit repozitářů DRAMBORA. Vývoje a testování se také účastnili pracovníci NK ČR, jejíž repozitář byl v roce 2007 auditu podroben s výsledkem 65 vážných rizik – více viz kapitola 6.4.3.1.

Právě díky projektu DPE začala problematika ochrany digitálních dat na půdě NK ČR dostávat pevnější obrysy. Projektové peníze z DPE, které NK ČR během projektu dostávala za odpracované člověkohodiny, byly investovány do personálního budování *Odboru digitální ochrany* a nabývání znalostí jeho pracovníků (návštěvy konferencí, knihoven apod.). Na základě zkušeností získaných během projektu DPE vznikl také plán na projekt NDK. Finance z DPE částečně umožnily první etapy přípravy projektu NDK. Z projektu DPE, zkušeností a kontaktů NK ČR čerpá dosud.

NK ČR se také účastnila v letech 2008-2010 dalšího projektu EU s názvem LiWa (*Living Web Archives*), který měl za cíl řešit problematiku dlouhodobé ochrany dat, tentokrát dokumentů získaných sklízením webu, které NK ČR v projektu *WebArchiv* provádí od roku 2000. Podobně v letech 2009-2010 se pracovníci *Odboru digitální ochrany* účastnili příprav obecných požadavků na LTP systém v rámci *Pracovní skupiny LTP*, jejímiž členy byly významné evropské národní knihovny (Nizozemí, Norsko, Španělska, Velké Británie, Německa a Švýcarska). Výstup podstatně pomohl přípravě projektu NDK.

3.2.4.3 Vývoj technického řešení – Long-term preservation (LTP) systém

Za jednu z prvních paměťových institucí, která realizovala potřebu specifikovat a vytvořit systém na logickou dlouhodobou ochranu digitálních dat (LTP systém), lze bezpochyby považovat Národní knihovnu Nizozemí (KB). Tato knihovna vytvářela od roku 1999 funkční specifikace LTP systému a stejného roku vypsala výběrové řízení na technické řešení. V uvedenou dobu nebyl na trhu žádný systém, který by měl funkcionalitu LTP systému nárokovanou KB. Výběrové řízení vyhrála firma

[65] http://www.casparpreserves.eu/
[66] http://www.planets-project.eu/
[67] http://www.casparpreserves.eu/training/Members/metaware/Events/wepreserve-dpe-planets-caspar-nestor-joint-training-event-starting-out-preserving-digital-objects-principles-and-practice.html
[68] http://www.digitalpreservationeurope.eu/publications/

IBM, která se pustila s KB do společného projektu. IBM deklarovala, že podobný systém nemá, ale že se pokusí ve spolupráci s KB jej vytvořit i přes to, že to bude velmi obtížné. To rozhodlo, ostatní firmy ve výběrovém řízení uvedly, že podobný systém vytvoří samy a bez problémů [RAS, 2009]. Od roku 2003 byl systém pod názvem e-Depot[69] plně funkční. Firma IBM spolu s KB vytvořila první LTP systém vůbec, který v rámci firmy IBM získal název DIAS (*Digital Information Archiving System*) a byl komerčně dostupný až do roku 2010. E-depot měl za cíl ukládat a ochraňovat publikace od nizozemských a později zahraničních vydavatelů[70] v rámci dobrovolného výtisku[71]. V roce 2011/2012 proběhne v KB nový tendr na LTP systém druhé generace. DIAS již neodpovídá funkcionalitě a projektům knihovny, která začala masově digitalizovat v roce 2010. DIAS nikdy neměl modul *Plánování ochrany*, jak je definován v referenčním rámci OAIS.
Systém DIAS byl implementován vedle KB také v německé národní knihovně, ovšem pouze v pilotním stavu, nikdy nebyl považován za ostrý provoz [ALTENHÖNER, 2009]. Knihovna chystá jeho náhradu od roku 2013. DIAS je považován za první generaci LTP systémů, kde byl jediný zástupce. Tento systém dnešním nárokům již nevyhovuje a není dále vyvíjen ani podporován.
Podobné cíle, a to zajistit dlouhodobou ochranu svých dat, měli i v Národní knihovně Nového Zélandu (NK NZ). Impulsem bylo schválení zákona o povinném výtisku, který začal platit od roku 2003. V tom okamžiku bylo jasné, že knihovna musí zaručit integritu, autenticitu a tím pádem důvěryhodnost uložených digitálních dat. V roce 2004 tak vznikl národní program NDHA (*The National Digital Heritage Archive*) k zajištění technického řešení a nákupu HW. Program je aktivní dodnes a NK NZ spolu s Národním archivem Nového Zélandu jsou lídry ve vývoji a výzkumu problematiky dlouhodobé ochrany digitálních dat. V rámci projektu vznikly mezi lety 2003-2005 funkční specifikace LTP systému, který dostal jméno KRONOS a byl vyvíjen od roku 2006 ve spolupráci s firmou Endevour, která dodávala knihovně automatizovaný knihovnický systém. Firma Endevour byla později koupena izraelskou firmou Ex Libris, spolupráce s NK NZ pokračovala nadále. Knihovna vytvářela a upravovala funkční specifikace a spolu s Ex Libris byl systém vyvíjen. LTP systém dostal pracovní název DPS (*Digital Preservation System*) a od verze 1 se jmenuje Rosetta. Systém se dostal do operační fáze v roce 2008. Dnes, po několika letech vývoje, je ve verzi 2.2 a jde o jeden z nejlepších LTP systémů, které jsou k dispozici[72]. Dubnu 2012 bude v Národní knihovně Nového Zélandu a Národním archivu Nového Zélandu nasazena verze 3. Vedle Rosetty je to dále LTP systém Safety Deposit Box od firmy Tessella z Velké Británie, který je rozšířen v archivech a několika knihovnách po celém světě. Oba systémy jsou funkčně v podstatě rovnocenné.

LTP systémy, nejen komerční, ale i institucionální nebo open source, lze dělit na dvě skupiny. Na LTP systémy první a druhé generace. První generace systémů vznikala na začátku nového tisíciletí. Instituce se tehdy snažily vybudovat systémy na logickou ochranu digitálních dat, které by odpovídaly referenčnímu rámci OAIS. Nejúspěšnější byla nizozemská národní knihovna s produktem IBM DIAS. Za podobný systém první generace můžeme považovat také DAITSS vyvíjený ve *Florida Digital Archive* používaný od roku 2005. Systémy LTP druhé generace jsou již výše zmíněná řešení (Rosetta, SDB, Archivematica), která si berou ponaučení z pionýrských

[69] http://www.kb.nl/hrd/dd/index-en.html
[70] Např. Kluwer, Elsevier, Springer, Blackwell, Oxford University Press, Taylor & Francis aj.
[71] Nizozemí nemá zákon o povinném výtisku a ten je nahrazován dohodami s vydavateli, kteří ochotně dokumenty zasílají.
[72] http://www.exlibrisgroup.com/category/RosettaOverview

začátků. Instituce, které se účastnily vývoje a měly LTP systémy první generace (národní knihovny Nizozemí, Německa, americké univerzity), začaly po roce 2010 cítit, že jejich řešení není již dostatečné a provedly nebo plánují provést přechod na systémy druhé generace. Ty mají šanci se rozšířit daleko více než systémy předchozí, zvláště proto, že se jedná o hotová lehce implementovatelná řešení. Bohužel, i podle průzkumu projektu PLANETS [SINCLAIR a BERNSTEIN, 2010], který proběhl s osmnácti ICT firmami, se ukazuje, že zájem firem je malý a trh s komerčními produkty určenými pro logickou dlouhodobou ochranu digitálních dat je stále v plenkách a to přesto, že poptávka po řešeních na ochranu digitálních dokumentů značně roste. V současné chvíli neexistuje ani náznak, že by vedle dvou výše jmenovaných komerčních systémů byl dostupný jiný nebo se alespoň vyvíjel.
I z toho důvodu se začaly od roku 2009 objevovat systémy na logickou dlouhodobou ochranu digitálních dat určené i běžným uživatelům a malým firmám. Např. Archivematica, HOPPLA, RODA nebo MOPSEUS.

3.2.1 Udržení autenticity, integrity digitálních objektů

S rozmachem používání a ukládání digitálních objektů se začalo ukazovat, že digitální objekty potřebují vyšší standard autenticity a integrity než tištěné informace. Autenticita a integrita jsou důležité pro dlouhodobé uchování dat a také pro důvěryhodnost repozitáře (to, co zpřístupňuje, je opravdu to, za co se vydává).
Integrita digitálního objektu (bistreamu) je ve velmi zjednodušeném pohledu dána neměnností bitů, ze kterých se skládá, případně částí, ze kterých se skládá komplexní digitální objekt. Integrita informace může ovšem spočívat také ve vlastnostech a doprovodných informacích k dokumentu se vztahujících. Zpráva *Preserving Digital Information* [GARRETT a WATERS, 1996, s. 11-18] uvádí, že integrita digitální informace má spojitost s obsahem, s celistvostí, referencemi, původem a kontextem.
Obsahem jsou myšleny formát a struktura, které obsah vyjadřují a dovolují jej zobrazit v konkrétní aplikaci či jinak použít. Právě formát a struktura bitů působí problémy z dlouhodobého hlediska pro různé systémy. Digitální repozitář nebo LTP systém musí překonat limity vyplývající z použití konkrétního HW nebo SW pro čtení a zobrazení digitálních dokumentů tím, že jsou známy přesně obsah a vlastnosti těchto objektů. Celistvostí je myšlena integrita jednoho konkrétního nespojitého (diskrétního) jednoduchého nebo komplexního digitálního objektu, která je nejčastěji fixována a následně validována pomocí tzv. kontrolních součtů. Reference souvisejí se skutečností, že digitální objekt musí mít pevnou možnost odkazování, která se nemění např. spolu s přesuny digitálního objektu. I to je jedna z věcí, které integritu dokumentu ovlivňují a dotvářejí. Aby byla zachována integrita objektu, jeho kompletnost, musí být možné jej kdykoliv během doby přesně a bez pochyb lokalizovat mezi ostatními objekty [GARRETT a WATERS, 1996, s. 15]. Původ (údaje o provenienci) jsou údaje o vzniku a životním cyklu digitálního objektu nebo dokumentu. Jde o prokázání všeho, co se s objektem dělo, kde a jak vznikl, tak, aby bylo možno na těchto údajích stavět procesy dlouhodobé ochrany. Koncový uživatel může v budoucnu požadovat údaje o všech změnách digitálního objektu, aby si byl jist, že se opravdu dívá na několikátou verzi originálního objektu, která se může lišit formátem i vlastnostmi, ale je autentická a s obsahem nebylo nedovoleně ani/nebo náhodně manipulováno. Poslední součástí, která spoludotváří integritu digitálního objektu nebo dokumentu, je kontext nebo kontextová informace. Kontext jsou vztahy mezi různými objekty nebo mezi objektem a prostředím (např. technickým i sociálním).

Poslední dvě hlediska, která ve své výše uvedené zprávě uvádějí Garrett a Waters, totiž informace o původu a kontextu, už z pohledu dnešních systémů dlouhodobé ochrany souvisejí spíše s autenticitou digitálního objektu a jejím vnímáním uživateli, producenty dat apod. Ovšem ani dnes není rozdíl mezi integritou a autenticitou digitálního objektu přijímán stejně.

Autenticita označuje nejčastěji v běžném světě skutečnost, že entita odpovídá stavu, který je pro ni výchozí nebo původní. Autenticita vždy porovnává stávající stav entity s něčím, co existovalo v minulosti. Pro digitální objekty se autenticita vztahuje nejčastěji k originálnímu objektu a jeho podobě, ve které vznikl nebo byl uložen v repozitáři. Jde o vztah objektu vytvořeného producentem a objektu, který je zpřístupněn uživateli. Digitální objekt musí doprovázet metadata, která umožní zpětně sledovat a kontrolovat všechny procesy, které se odehrály mezi dobou, kdy byl digitální objekt vytvořen, a okamžikem, kdy si jej prohlíží nebo jinak používá uživatel. Z pohledu digitálního objektu a logické dlouhodobé ochrany digitálních dat je autenticita jistota, že objekt předkládaný uživateli nebyl změněn, je v původní formě a nikdo s ním nemanipuloval; nebo je změněn (prošel například migrací formátu), ale o všech změnách je záznam v metadatech a samotný obsah objektu (dokumentu) je shodný s původním obsahem, tedy nebyl měněn. Každé rozhodnutí o autenticitě, zda je nebo není, je založeno na porovnání stávajícího stavu se stavem předchozím a to na základě integrity a odpovídajících údajích o původu (provenienci) [GLADNEY, 2007, s. 106]. Z výše uvedeného je jasné, že pro zachování a prokázání autenticity i integrity je nutné mít každý digitální objekt uchován spolu s důkazy autenticity a integrity ve formě záznamů metadat různých typů.

3.2.2 Signifikantní vlastnosti digitálních objektů a metadata

Dlouhodobá ochrana digitálních dat je o uchování vlastností nejrůznějších digitálních objektů skrz procesy ochrany, kterými jsou např. migrace nebo emulace. Tento proces je ale většinou takový, že ne všechny vlastnosti objektů jsou zachovány. Kvalita digitálního objektu musí být během životního cyklu kontrolována. Je nutno si stanovit, které vlastnosti jsou podstatné a které lze oželet. Podstatné se označují jako signifikantní (angl. *significant properties*). Ztráta vlastností nemusí proběhnout pouze po migraci, ale také díky vývoji technologií a tedy např. změnám SW aplikací, které dostatečně nepodporují předchozí verze.
Pro všechny digitální objekty je nutně potřeba jejich vlastnosti dobře zdokumentovat, nejlépe v podobě metadat, která jsou uložena spolu s digitálním objektem. Signifikantní i jiné vlastnosti musí být k digitálnímu objektu doplněny nejpozději v okamžiku vstupu do digitálního repozitáře. Ideální ovšem je, když tyto údaje provázejí objekt od okamžiku jeho vzniku. Z pohledu dlouhodobé ochrany jsou tyto údaje klíčové pro další zacházení s objektem, jeho správu a případné opatření dlouhodobé ochrany (migrace). Právě vůči těmto vlastnostem lze, a je nutné, provést srovnání výsledku migrace k původnímu objektu. Musíme být schopni garantovat, že konkrétní nová verze digitálního objektu je ekvivalentní ve svých signifikantních vlastnostech k objektu původnímu, i přesto, že je např. na zcela jiné technologické platformě. Jedna z definic říká, že signifikantní vlastnosti jsou: „*Charakteristiky digitálních objektů, které musí být chráněny během doby, aby zajistily kontinuální přístupnost, použitelnost a smysl objektů.*" [WILSON, 2007]

Podle toho, které vlastnosti chceme zachovat, musíme vytvářet strategie ochrany, volit nástroje a specifikovat metadata a vlastně i datové formáty. Jde o to určit, jaké vlastnosti jsou pro nás

důležité, budeme je zapisovat do metadat a uchovávat tak pro budoucnost. Na základě této úvahy lze vybrat vhodné metadatové schéma. Signifikantní vlastnosti jsou spojeny s autenticitou, protože tu je možno posoudit právě díky zachování a popisu vlastností konkrétního digitálního objektu. Signifikantní vlastnosti jsou dané potřebami instituce, cílem uložení dat a také zájmy cílové komunity, tedy uživatelů. U různých institucí se tedy často liší. Nejčastěji se vytvářejí přehledy signifikantních vlastností pro konkrétní datové formáty. Z technických vlastností se uvádějí a sledují mj.:

- obsah – digitální objekt je textový, obrazový,
- kontext – kdo, kdy a kde digitální objekt vytvořil,
- vzhled – font, velikost písma, barva, celkový tvar,
- chování – hyperlinky, vzorce v MS Excel aj.,
- struktura – vložené objekty, stránkování, nadpisy.

Obecné signifikantní vlastnosti nejsou technického rázu, ale v podstatě popisují cíle, které logická dlouhodobá ochrana digitálních dat sleduje. Margarita Korenkova a Ann Hägerfors ve svém článku *Quality Criteria for Digital Information in Long-term Digital Preservation* [KORENKOVA a HÄGERFORS, 2011], vedle skvělého rozboru existující literatury k signifikantním vlastnostem, tyto obecné vlastnosti uvádějí. Mj. jmenují: přesnost (digitální objekt je spolehlivý, neobsahuje chybu); autenticita; kompletnost; kontextovost; identifikovatelnost (objekt lze nalézt); interpretovatelnost (objekt lze použít a pochopit); čitelnost; relevantnost; důvěryhodnost aj.

Základní sady vlastností digitálních objektů jsou v knihovnické komunitě do jisté míry dány a podporovány registry (jako je např. PRONOM nebo GDFR), které obsahují specifikace a výčty vlastností jednotlivých formátů. Schéma PREMIS specifikuje mnoho elementů metadat, která popisují vlastnosti objektů a má i speciální část pro signifikantní vlastnosti. Jde o element <significantProperties>, který patří k popisu entity Object. Obsahuje vnořené elementy <significantPropertiesType>, <significantPropertiesValue> a <significantPropertiesExtension>. Tímto způsobem, tedy pomocí typů a hodnot, lze vyjádřit jakékoliv signifikantní vlastnosti v PREMIS a ještě případně připojit XML, které tyto údaje rozšiřuje. PREMIS pojímá signifikantní vlastnosti obecně pro jakýkoliv digitální objekt a dokáže je také vyjádřit. Naproti tomu např. registr formátů PRONOM se dívá na signifikantní vlastnosti z pohledu jednotlivých formátů, u kterých je shromažďuje do databáze.

3.2.3 Možná řešení a opatření dlouhodobé ochrany digitálních dat

Digitální objekty jsou nejčastěji ohroženy ze dvou důvodů. Prvním z nich je nestálost média, na kterém jsou uloženy, druhým důvodem je závislost jejich použitelnosti a zobrazitelnosti na SW aplikaci. Strategie dlouhodobé ochrany digitálních dat můžeme rozdělit do několika skupin: 1) ochrana technologií (počítačové muzeum, ochrana HW); 2) emulace technologií (tvorba emulačního SW, případně digitální archeologie); 3) migrace informací (převod datových formátů a normalizace) a 4) ostatní (zapouzdření, přenos dat na papír). Existující přístupy a metody jsou uvedeny ve stručnosti níže:

Formátová migrace – změna digitálních objektů, např. závislých na jedné technologii (HW a SW), na jiný formát, použitelný na nových technologiích při zachování obsahu digitálních objektů, použitelnosti, zobrazitelnosti. Na rozdíl od obnovování, které je pouze udržováním bitstreamu

(např. přesun dat na nový HW beze změny formátu dat), dojde při formátové migraci ke změně digitálního objektu, tedy původního bitstreamu.[73]

Obnovování nosiče dat/fyzická migrace – není metodou logické ochrany dat, ale ochrany základní, tedy bitstreamové. Jde o náhradu starých médií za nová shodná nebo za nová založená na nové technologii (např. přenos dat z optických disků na magnetické LTO pásky). Digitální objekt jako takový zůstává nedotčen a neměněn. Jediným problémem je zachování celistvosti a integrity. Ke kontrole integrity bitstreamu se používají kontrolní součty, což jsou algoritmy, které využívají bitstream digitálního objektu pro vlastní výpočet. Výsledkem je řetězec znaků, který může být uložen jako další bitstream, a který se nazývá kontrolní součet. Tento kontrolní součet lze zpětně použít ke kontrole, zda se na původním bitstreamu něco změnilo. Nejznámější metody jsou SHA-1 a MD5.

Normalizace/spoléhání na standardy – do archivu přicházející digitální objekty se převádějí do omezeného počtu preferovaných formátů, u kterých se předpokládá, že splňují nároky na dlouhodobé uložení. Takovými nároky může být dostupná dokumentace k datovému formátu, jeho rozšířenost, dostatek vhodných aplikací, robustnost, odolnost vůči poškození apod. Nelze předpokládat, že konkrétní datový formát vydrží věčně, je vždy nahrazen. Jde jen o to odhadnout vhodnost a dlouhodobost využívání konkrétního formátu, což může ve výsledku ušetřit práci s migrací nevhodného formátu a také finance. Normalizace u většiny digitálních úložišť je aplikována, většina z nich má seznam preferovaných formátů. Spolu s novým (normalizovaným) digitálním objektem se musí uložit i původní objekt, pro budoucí použití a prokázání autenticity nebo další migrace do datových formátů v současné době neznámých. Menší počet známých a ověřených formátů je méně náročný na správu v rámci repozitáře. Ovšem pozor, repozitář musí být schopen přijmout a uložit jakýkoliv formát, pokud to bude nutné. Nesmí být omezen pouze pro formáty preferované.[74]

Emulace – proces vytváření aplikací napodobujících zastaralé systémy a technologické platformy na současných nebo budoucích počítačích pomocí emulace SW, HW a operačních systémů. Digitální objekt v původní podobě tak lze použít i na novém HW a SW.

Digitální archeologie – nasazení procesů k obnovení a zpřístupnění digitálních objektů ze zastaralých nebo poškozených fyzických médií. Přístup je tedy takový, že se s daty nebude nic dělat a problém se začne řešit, až bude v budoucnu data někdo chtít použít. Problémem je, že data mohou být obnovena, ale nebude obnoven neexistující kontext a je možné, že data nebudou bez odpovídajících metadat dávat žádný smysl. Nemají žádné informace o reprezentaci ve smyslu OAIS referenčního rámce. Obnova také může trvat velmi dlouho, pokud bude vůbec úspěšná.

---

[73] V problematice dlouhodobé ochrany digitálních dat se často mluví o uchování originálu. Je však těžké určit, co je oním originálem. Ve fyzickém světě je to jednoduché, ve světě digitálním lze těžko o originálu mluvit, snad jen o prvotní manifestaci informačního zdroje. Originál pro digitální repozitář je to, co do repozitáře původně bylo uloženo a bylo tedy schváleno, zkontrolováno a považováno za dokument, který byl ve všech směrech v pořádku.

[74] V českém prostředí má seznam preferovaných formátů např. oblast archivnictví v podobě vyhlášky Ministerstva vnitra č. 191/2009 Sb. o podrobnostech výkonu spisové služby. Tato vyhláška specifikuje, jaké typy datových formátů digitálních objektů různých typů jsou výstupem z tzv. elektronických systémů spisové služby. Tento výstup je pak vstupem do archivního systému (LTP systém Národního archivu, který by měl být v provozu od roku 2012). Vyhláška specifikuje formáty pro obrazové dokumenty (PNG, JPG, TIFF), textové dokumenty (PDF/A), dynamické dokumenty (MPEG-2, MPEG-1 a GIF), zvukové dokumenty (MP2, MP3, WAV, PCM) [ČESKO. MINISTERSTVO VNITRA, 2009, s. 2781]. Metadata musejí odpovídat specifikaci NSESS (Národní standard pro elektronické systémy spisové služby).

HW muzeum/uchování technologií – spočívá ve shromažďování různých HW z různých technologických etap vývoje počítačů, spolu s odpovídajícím SW. Tento přístup se nedoporučuje v dlouhodobém horizontu. Díky omezené životnosti technických přístrojů jakéhokoliv typu i v HW muzeu v určitém okamžiku nastane situace, kdy se rozbije poslední funkční exemplář konkrétního typu počítače, jeho součásti nebo periferie. Navíc tento přístup navíc předpokládá, že se digitální objekty zachovají pouze na původních nosičích, což je spíše výjimka.[75] Přístup HW muzea nepočítá ani s náročností na obsluhu starých přístrojů. Možnost zpřístupnění na několika málo přístrojích schopných zobrazit konkrétní dokument je navíc velmi omezená.

Přenos dat na klasický nosič – papír nebo mikrofilm. Nejde o vhodné řešení, mnoho typů digitálních dokumentů není možné vytisknout tak, aby neztratily smysl, existující vazby nebo některé ze svých dynamických vlastností (např. webové stránky apod.). Hlavní vlastností, která by se ztratila tímto opatřením, je strojová čitelnost, a to u všech typů dokumentů, včetně těch textových. Pokud se k opatření přistoupí, měl by se použít permanentní papír, který má určitý obsah zásaditých složek dle norem ISO 9706:1994 navazující na obdobný americký standard ANSI/NISO Z39.48-1984. Přenos na mikrofilmy není tak bezpředmětný, jak by se mohlo zdát. Existují metody, jak zachytit na mikrofilm digitální informaci pomocí kódu ve speciální SW aplikaci. Ke čtení takového digitálního dokumentu uloženého na mikrofilmu je nutné informaci dekódovat opět pomocí konkrétního SW. Výhodou je skutečnost, že mikrofilm jako takový vydrží nejméně pět set let bez zvláštních nároků na uložení a bez zvláštních nároků na financování, oproti klasickému uložení na HW. Nevýhodou je nutnost použít na dekódování informací SW.

Zapouzdření – zabalení digitálního objektu spolu s metadaty, identifikátory a s prvky nutnými pro pozdější přístup k němu. Kontejner tak obsahuje detaily o tom, jak má interpretovat bitstream digitálního objektu v kontejneru uloženém. Může také obsahovat informace o nárocích na emulaci, může obsahovat i SW aplikaci pro zpřístupnění, operační systém a další dokumentaci. Podobný přístup popsala v roce 2002 Abby Smith ve svém článku *Rethinking the Library in Digital Age* pod názvem ***Persistent Object Preservation*** (POP) [SMITH, 2002, s. 10]. Její koncept spočíval v tom, že jsou jasně deklarovány vlastnosti původní digitální informace (např. obsah, struktura, kontext), které zaručují odolnost. Tento přístup je jediným přístupem ochrany digitálních dat, který se zaměřuje na ochranu digitální informace už při jejím vzniku. Ostatní strategie ochrany se pokoušejí „přelstít“ technický problém zastarávání „ex-post“.

Jednotlivé možnosti řešení dlouhodobé ochrany navzájem nesoupeří, ale mohou se v jedné knihovně nebo v projektu dobře doplňovat. Dosti často data z jednoho digitálního repozitáře mohou být natolik různá, že budou aplikovány dva přístupy. Např. pro digitalizovaná data v TIFFu migrace, pro data z archivace webu to může být emulace. Výběr způsobu a metody je nutné rozmyslet dopředu, upravit a vytvořit dle vybrané metody metadata a také datové formáty. Nutné je promyslet proveditelnost, náklady finanční a časovou náročnost. Podstatou jakéhokoliv řešení musí být snaha překonat problémy spojené se zastaráváním technologií a to jak z technického hlediska, tak z hlediska administrativního. Toto překonání nesmí mít za následek nechtěnou změnu digitálního objektu. Může dojít k určitým změnám, ale ty musí být „chtěné“ a dokumentované (změna vlastností dokumentu apod.). Klíčová jsou v tomto směru metadata, která poskytují kontext a technické i administrativní informace o vlastních datech.

[75] Starý soubor např. ve formátu .t602 se nedochová na 8 palcové disketě, ani na 3,5 palcové, ale nejspíše na moderním nosiči, jako je CD nebo flash disk, kam se dostane např. kopírováním z média na médium. Pak je zpřístupnění pomocí HW muzea bezpředmětné a v podstatě nemožné.

Možnosti ochrany digitálních dat lze také vnímat z krátkodobého, střednědobého nebo dlouhodobého hlediska. Rozdělení samozřejmě není pevně dané a může se v konkrétních případech lišit nebo překrývat. Do krátkodobých přístupů řadíme technologické muzeum, obnovování dat. Do střednědobých pak emulaci a migraci a do opravdu dlouhodobých přístupů např. digitální archeologii nebo převod do analogové formy, ač jsou z dnešního pohledu nepřijatelné. Migrace i emulace jsou na hranici střednědobé a dlouhodobé strategie. Jde o dvě nejčastěji prováděné metody, které jsou blíže představeny v kapitole 3.2.4.

3.2.4 Emulace a migrace

Konceptuálně je ochrana digitálních dokumentů naprosto odlišná od ochrany klasických (např. papírových) dokumentů. Papírový dokument je uložen v odpovídajících klimatických podmínkách a cílem je, aby se nijak neměnil. U digitálních objektů je cíl stejný, vlastnosti, vzhled a pocit z dokumentu by se měnit neměly, ale cesta, jak toho docílit, je jiná. K tomu, aby digitální dokument byl ve svých vlastnostech stejný a přístupný i za pár desítek let, je nutné jej měnit (migrovat) a tyto změny pečlivě dokumentovat. Změny vlastností jsou možné pouze tehdy, pokud o nich víme a jako správci dokumentů jsme ochotni je akceptovat (problematika tzv. signifikantních vlastností).
Emulace a migrace jsou nejčastěji využívanými aktivitami ochrany, jak je definuje referenční rámec OAIS.[76]

3.2.4.1 Migrace

Migraci lze obecně popsat jako přenos digitálních objektů ze starého technologického prostředí do prostředí nového. Cílem je zachování možnosti zpřístupnění digitálních objektů i v nových technologických prostředích. To se může týkat jak HW (přenos na nový repozitář – fyzická migrace), tak SW (nutnost změny datového formátu – formátová migrace). Průlomová zpráva *Preserving Digital Information* [GARRETT a WATERS, 1996] popisuje, že migrace v dlouhodobé ochraně digitálních dat je více než pouhé obnovení dat nebo přenos dat. Autoři uvádějí, že: *„migrace je sada organizovaných úkonů přizpůsobených k dosažení periodického přenosu digitálních materiálů z jedné HW/SW konfigurace na druhou, nebo z jedné generace počítačové technologie na novou generaci. Cílem migrace je ochránit integritu digitálních objektů a zachovat pro uživatele možnost vyhledávat, zobrazovat a jinak je používat i přes neustálé změny technologií. Migrace je pro digitální archivy klíčová.“* [GARRETT a WATERS, 1996, s. iii] Podrobnější definici nabízejí Lawrence a kol. ve zprávě *Risk Management of Digital Information: A File Format Investigation* [LAWRENCE, et al., 2000, s. 2]: *„… migrace mění strukturu originálního datového souboru. S výjimkou souborů s velmi jednoduchým data streamem, většina souborů obsahuje dvě základní části: strukturální prvky a datové prvky. Formát souboru reprezentuje uspořádání strukturálních a datových prvků jedinečným a specifickým způsobem. V tomto smyslu je migrace proces přeuspořádání originálních sekvencí strukturálních a datových prvků (zdrojový formát) tak, aby odpovídaly jinému uspořádání (cílový formát)“*.

[76] Ochrannými aktivitami ale mohou být také např. uchování původního bitstreamu po normalizaci během vstupu dat do repozitáře; přidávání metadat k digitálním objektům apod. Tyto jsou do jisté míry předpokladem k pozdější migraci nebo emulaci.

Velkou výhodou migrace je, že jde o léty prověřený koncept, který je součástí dění okolo informačních technologií od jejich počátku. Migrace se neprovádí pouze pro záchranu přístupnosti archivních dokumentů, ale může mít mnoho jiných důvodů a cílů. Např. vytváření uživatelských kopií (TIFF>JPEG 2000), normalizace na vstupu do archivu (např. MS Word > OpenOffice) apod. Další z výhod je, že digitální objekt je uložen v podobě, která umožní jeho okamžité využití a není třeba udržovat staré HW a SW prostředí pro tuto příležitost. Migrace je vzhledem k technologickým změnám a pokroku nevyhnutelný proces v případě, že je zvolena jako jedna z metod dlouhodobé ochrany. V průběhu času musí probíhat opakovaně, vždy se budou objevovat nové a nové datové formáty, do kterých bude nutno původní digitální objekty převádět. Migrace tak reaguje na vývoj technologií. Pokud máme dobré znalosti o formátu, do kterého se migruje a je vhodně vybrán, může to oddálit další nutnou migraci a celý proces zlevnit i zjednodušit.
Problematickou stránkou migrace zůstává možná ztráta informací a signifikantních vlastností (např. tzv. *„look and feel"*) digitálního objektu. Úroveň a podoba ztráty vlastností závisí na tom, jak si instituce nastaví zachování signifikantních vlastností, jež chce při migraci zachovat. Nemusí např. chtít zachovat možnost jednoduché editace objektu a pak migrace z formátu MS Word do formátu PDF/A nemusí představovat problém, i když určitá vlastnost dokumentu je migrací ztracena. Za nevýhodu migrace je za určitých okolností možno považovat i to, že je nutné přesně specifikovat signifikantní vlastnosti všech typů objektů nebo formátů, a to před samotnou migrací. Po ní je nutné zkontrolovat, zda tyto vlastnosti objekt neztratil. Jinými slovy, migrace nárokuje neustálé plánování ochrany a to pro každý formát a každou jeho verzi zvlášť. Migrace komplexních dokumentů není zdaleka vyřešena a mnohdy se spoléhá na emulaci, jako v případě dat z archivace webu.
Často se také zmiňuje nejistota výsledku migrace. Tady ovšem již existují nástroje, které umožňují testovat jednotlivé SW nástroje na migrace, porovnat dle různých hledisek a signifikantních vlastností výsledky testovacích migrací (nástroj PLATO a PLANETS Testbed). Migrace mnoha milionů digitálních objektů může trvat velmi dlouho, měsíce i roky. Toto provádějí v rámci modulu Plánování ochrany i komerční LTP systémy Rosetta a SDB.

Je důležité si uvědomit, že migrace může probíhat v různých okamžicích životního cyklu digitálního dokumentu:

- při vstupu do repozitáře, kdy se mnohdy digitální objekty převádějí na vhodnější a preferované formáty,
- na vyžádání při konkrétním riziku jako hromadná ochranná aktivita (např. LTP systém Rosetta navrhuje migrace až v případě zastarání konkrétního formátu, které dává do souvislosti s tím, že neexistuje aplikace vhodná k zobrazení objektu[77]),
- pokud si uživatel vyžádá dokument, který je v zastaralém formátu.

Migrace také může probíhat nepozorovaně, např. při otvírání dokumentu vytvořeného ve starém MS Word v jedné z jeho novějších verzí, případně v OpenOffice aplikaci.[78]
Jeff Rothenberg, který je považován za odpůrce migrace, ve své práci *Avoiding technological quicksand* [ROTHENBERG, 1999, s. 14] uváděl, že problémem migrace je nemožnost určit, kdy je potřeba ji u konkrétních dokumentů provést, díky nemožnosti předjímat směr vývoje technologií.

[77] Tato okolnost může nastat i při vkládání objektu do repozitáře, v takovém případě je změněn formát a uložen nový formát i s původním objektem.
[78] Tato vlastnost určitých SW balíků je někdy využívána, představují vlastně nástroj na migraci, který lze zapojit do automatizovaného workflow. Běžně se takto využívá balík OpenOffice.

Dnes již ovšem existují systémy (LTP systémy), které na základě metadat a díky spolupráci s externími službami, jako jsou registry formátů (PRONOM, UDFR), jsou schopny správci repozitáře sdělit, že konkrétní formát začíná zastarávat a nabídnout možnosti migrace na formát jiný. A to včetně možností testování výsledků migrace a jejich hodnocení. Digitální objekty v ideálním případě obsahují maximální množství metadat již při svém vstupu do repozitáře (tj. poskytne je původce digitálního objektu). Existují i nástroje, které automaticky z digitálních objektů při jejich vstupu do repozitáře potřebná metadata vytvoří (většinou technická metadata). Jedním z nástrojů je např. známý JHOVE1 a verze 2[79]. Obsahují údaje o formátu, jeho verzi aj., se kterými pak LTP systém pracuje.
Z důvodů bezpečnosti, autenticity a možnosti návratu ke konkrétní verzi formátu, si LTP systém vždy nechává původní digitální objekt a pak také buď všechny následné verze vzniklé migrací, nebo jen ty verze, které znamenají zásadnější změnu signifikantních vlastností digitálního objektu. Právě díky metadatům (technickým a administrativním) jsou LTP systém schopny tuto funkcionalitu nabízet a tím podstatně zlehčovat práci administrátorům, zvláště v případech, kdy v úložišti jsou uloženy miliony objektů ve více datových formátech a navíc v různých verzích těchto formátů. LTP systém může napomáhat i s načasováním migrací, jejich postupným průběhem a jejich sledováním.

3.2.4.2 Emulace

Pod pojmem emulace se rozumí umělé „napodobení“ softwarového a hardwarového prostředí, které bylo typické nebo nutné pro prohlížení digitálního objektu v době jeho vzniku. Jde o jednu z metod logické dlouhodobé ochrany, která zatím stojí ve stínu migrace. Emulace může existovat na třech úrovních: na úrovni SW aplikace; na úrovni systémového SW (operační systém); a na úrovni hardwarové. V emulovaném operačním systému na emulovaném HW lze také přímo použít originální SW aplikace, pokud jsou k dispozici. Obě úrovně emulace SW jsou často proprietární, bez dostupné dokumentace a jejich emulace je proto velmi náročná. K tomu, abychom emulovali SW aplikaci nebo operační systém odpovídajícím způsobem, je nutné velmi dobře znát jejich technické vlastnosti, design a podmínky nasazení. Velmi dobrý popis emulace, jejích typů a možností, včetně typů emulátorů podávají J. van der Hoeven a H. van Wijngaarden v článku *Modular emulation as a long-term preservation strategy for digital objects* [HOEVEN a WIJNGAARDEN, 2005].

Výzkum emulace jako jedné z metod logické dlouhodobé ochrany digitálních dat začal po roce 1996 [GARRETT a WATERS, 1996]. Za nejvýznamnějšího propagátora emulace je dodnes považován Jeff Rothenberg, který ve své práci *Avoiding technological quicksand: finding a viable technical foundation for digital preservation* [ROTHENBERG, 1999] popsal zásady této metody a vyzdvihl ji nad metody ostatní (včetně migrace). Uvádí, že nejjednodušší metodou dlouhodobé ochrany digitálních dat je jejich uložení v původní podobě spolu se SW nutným pro jeho zobrazení/použití, samozřejmě spolu s metadaty popisujícími jak digitální objekty, tak i digitální prostředí, ve kterém vznikly, a ve kterém je možné je zobrazit/použít [ROTHENBERG, 1999, s. vi]. Aby tento přístup byl funkční, je nutné připojit i dokumentaci k SW jak technickou, tak uživatelskou. Používat původní aplikaci za 50 let totiž bude pro většinu uživatelů nepřekonatelný problém, včetně dohledání dokumentace. Koncepčním problémem, často emulaci vytýkaným, je

[79] http://hul.harvard.edu/jhove/

skutečnost, že se zabývá uchováním funkcionality systémů nutných pro zpřístupnění digitálních objektů, namísto toho, aby se zabývala uchováním samotných digitálních objektů – viz [BEARMAN, 1999], který Rothenbergovu analýzu ve svém článku ostře napadl, převážně díky jeho odsuzování migrace jako nefunkční metody dlouhodobé ochrany. Právě migrace, argumentuje Bearman, je jedinou dnes funkční metodou. Podobně viz [GRANGER, 2000].
Evropským lídrem ve výzkumu emulace se stala Národní knihovna Nizozemí, díky spolupráci s IBM na projektu e-depot, díky pozdějšímu vývoji emulátoru DIOSCURI[80] a zapojení v evropském projektu KEEP[81] (*Keeping Emulation Environments Portable*; FP7 2009-2012). Projekt se zabývá emulací HW i SW prostředí pro digitální objekty. Cílem je vytvořit strategii pro poskytnutí permanentního zpřístupnění dynamického multimediálního obsahu (hudba, multimédia, webové stránky, databáze i počítačové hry) v dlouhodobém horizontu [HOEVEN, SEPETJAN a DINDORF, 2010, s. 113]. Ze starších projektů, které se zabývaly emulací, jmenujme evropský projekt NEDLIB (1998-2000), který se mj. zabýval emulací jako strategií dlouhodobé ochrany elektronických časopisů – více viz [VAN DER WERF-DAVELAAR, 1999]; a britský projekt CAMiLEON (1998-2003), který řešil znovuzpřístupnění tzv. *Domesday Book* v projektu BBC – viz kapitola 3.2.

Nepopíratelnou výhodou emulace je, že můžeme mít digitální objekt v repozitáři uložený jen a pouze v jeho původní podobě. Nemusíme ho jednou za čas migrovat do stále nových a nových formátů a hromadit tak jeho jednotlivé verze. Emulace také může být u spousty typů komplexních digitálních objektů jedinou metodou, jak je zobrazit a zpřístupnit v budoucnu (webové stránky[82], databáze). Nevýhodou naopak je, že emulační software stále není vyvinut natolik, aby tento přístup poskytoval záruku, že digitální objekty opravdu bude možné kdykoliv zpřístupnit v jejich původním prostředí [HUTAŘ, 2008c, s. 10]. Navíc emulační software čelí stejnému problému jako digitální objekty samotné, v horizontu několika let bude sám potřebovat další emulátor, aby bylo možné původní emulátor spustit. Emulátory a emulovaná prostředí představují tak komplikované soustavy vztahů a závislostí, že je často problém udržovat v chodu je samotné. Komplikací může být i to, že k úspěšné emulaci prostředí pro konkrétní digitální objekt je potřeba mít jasnou představu o prostředí, ve kterém objekt vznikl a pro které byl určen. Tyto údaje musejí být součástí tzv. ochranných metadat (např. PREMIS). Ve spolupráci s projektem KEEP vznikla aktivita TOTEM[83] (*Trustworthy Online Technical Environment Metadata*). V projektu TOTEM vzniká databáze určená k zaznamenání údajů o určitých kombinacích HW a SW pro konkrétní formát digitálního objektu. Vzniklo také metadatové schéma TOTEM, které tyto kombinace popisuje. Databáze obsahuje údaje o verzích SW, potřebné dynamické knihovny (DLL), operační systémy včetně updatů a také HW komponenty, včetně verzí. Jde o prostředí, kde lze ukládat výše zmíněné informace a také v nich vyhledávat, např. podle typu formátu, operačního systému apod. Schéma TOTEM bude součástí nové verze ochranných metadat PREMIS.
Schopnost přesně reprodukovat digitální objekt závisí na schopnostech emulátoru. V současnosti je možné „napodobit“ mnoho kombinací SW a HW, ale zdaleka ne všechny. Mnohé SW mohou

[80] http://dioscuri.sourceforge.net/
[81] http://www.keep-project.eu/
[82] Obsahují extrémně velké množství typů datových formátů, u kterých není možné na vstupu do archivu některé přijímat a jiné odmítat. Principiálně je metoda migrace možná, jak ukazují např. LTP systémy SDB od firmy Tessella a Rosetta od firmy Ex Libris. Jde jen o rozhodnutí, jak se bude k archivaci dat z archivace webu přistupovat a jak se budou zpřístupňovat.
[83] http://keep-totem.co.uk

obsahovat ochranu proti kopírování nebo mechanismy aktivace, které v emulovaném prostředí nelze provést. Mohou tak limitovat nebo zcela zabránit svému použití. Jak se ukázalo, tak právní problémy jsou v emulaci velmi tvrdým oříškem. O tomto typu problému mluvil už v roce 2000 Stewart Granger [GRANGER, 2000]. Právní implikace emulace byly zkoumány a analyzovány v projektu KEEP v letech 2009 a 2010 a jsou popsány mj. v článku *Legal aspects of emulation* [HOEVEN, SEPETJAN a DINDORF, 2010]. Pochybnosti o protiprávním zacházení se týkají aktivit, jako jsou přenos dat z médií chráněných proti kopírování (tedy obcházení tohoto opatření) na druhá média; použití a úpravy SW zatíženého autorskými právy; emulace patentovaných komponentů HW řešení.

Emulace tedy, i přes naděje do ní vkládané, není tak přímočarou metodou a často naráží na netušené problémy. I proto je migrace stále silně preferovaná na úkor emulace. Potvrdil to i průzkum ICT firem a řešení, které nabízejí v této oblasti, který proběhl v roce 2009 [SINCLAIR a BERNSTEIN, 2010, s. 3]. Od roku 2010 se emulace ale stále častěji prosazuje jako použitelné řešení a výzkum se rozšiřuje. Vznikají první reálná řešení a implementace – viz prezentace na konferenci iPRES 2011 v Singapuru, kde téma emulace zcela převládalo [BORBINHA, et al., 2011].

3.2.5 Nástroje a služby třetích stran využívané pro dlouhodobou ochranu digitálních dat

V komunitě okolo dlouhodobé ochrany digitálních dat je od počátku velká poptávka po nástrojích, které by v určitých okamžicích životního cyklu digitálního objektu automaticky prováděly konkrétní úkoly. Mezi tyto úkoly lze zařadit charakterizaci, ochranu dat (plánování ochrany, migrace, emulace), manipulaci s daty, správu dat apod. První nástroje začaly vznikat po roce 2000, jejich použitelnost a počet se od té doby zvyšuje. V textu níže jsou uvedeny ty nejpodstatnější, které se váží jak k dlouhodobé ochraně, tak k metadatům. Vynechány jsou ostatní nástroje a metody např. na přenos dat (BagIt), konkrétní nástroje na migrace (např. ImageMagick), LTP systémy aj.

3.2.5.1 Nástroje na charakterizaci

Nejpoužívanější jsou nástroje na charakterizaci digitálních objektů. Jde o proces získávání informací z digitálního objektu. Tyto informace jsou vytvořeny a poté uloženy ve formě metadat. Proces charakterizace má tyto části:

- *identifikace formátu* – proces ověření předpokládaného formátu digitálního objektu na základě externích příznaků (přípona názvu souboru) a vnitřních charakteristických rysů, tzv. podpisů (*format signatures*). Identifikace pomocí podpisů funguje tak, že nástroj hledá typické struktury v kódu, který tvoří soubor. Struktury mohou být jednoduché řetězce znaků nebo více řetězců v určitém pořadí na různých místech kódu souboru. Každý formát má tyto struktury odlišné. Jednoduchá otázka by zněla: je soubor s příponou PDF opravdu PDF, za které se vydává?
  - Nejčastěji používané nástroje jsou DROID (s pomocí databáze PRONOM), JHOVE1 a JHOVE2, FIDO.
- *validace formátu* – proces ověření do jaké úrovně odpovídá digitální objekt syntaktickým a sémantickým pravidlům definovaných specifikací formátu objektu. Nástroj validace

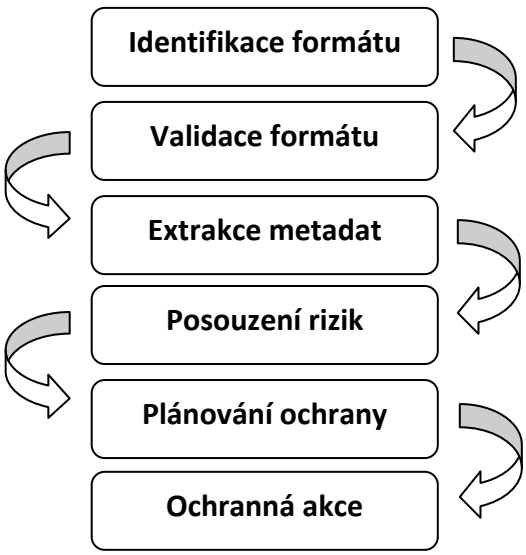
prochází celý soubor, identifikuje a mapuje každou jeho sekci, porovnává míru, v jaké odpovídají specifikaci. Nakonec vytvoří výsledek, který může být podrobný, nebo jde pouze o konstatování, že soubor specifikaci, ke které se hlásí, odpovídá nebo ne. Jednoduchá otázka by zněla: je PDF validní podle specifikace?

  - Nejčastěji využívané nástroje jsou XCEL, JHOVE1 a JHOVE2.

- *extrakce vlastností objektu (metadat)* – proces získání a zaznamenání údajů o vlastnostech digitálního objektu signifikantních pro opatření dlouhodobé ochrany (tedy převážně technická metadata). Výstupem je XML záznam.
  - Nejčastěji používané nástroje jsou JHOVE1 a JHOVE2, FITS, Kakadu, NZME.

Identifikace a validace formátů i extrakce metadat jsou nutné pro procesy dlouhodobé ochrany i pro správu digitálních objektů v digitálních repozitářích. Rozhodnutí o uložení, změnách, ochraně jsou založena nejčastěji na formátu digitálního objektu, který tedy musí být správně určen a uchován v metadatech. Všechny procesy validace, identifikace apod. musí probíhat automatizovanou cestou. K tomu jsou využívány nástroje jako je JHOVE, PRONOM/DROID, které z pohledu OAIS vytvářejí a do metadat ukládají tzv. informaci o reprezentaci. Jsou hlavním zdrojem technických i jiných metadat pro digitální objekty určené k dlouhodobé ochraně. Výstupy těchto nástrojů se ukládají buď celé (např. v rámci METS nebo PREMIS), nebo jsou mapovány do dalších metadatových schémat (MIX a PREMIS). Některé nástroje jsou schopné výstupy dodat přímo ve standardních schématech. Bez uvedených nástrojů se žádná aktivita dlouhodobé ochrany nemůže obejít. Existuje několik seznamů relevantních nástrojů, zmínit bychom měli alespoň aktivitu *Open Planets Foundation* a její registr[84] a také seznam knihovny Kongresu[85].

Celý proces charakterizace probíhá při vstupu digitálního objektu do repozitáře (nebo LTP systému), případně již v procesu digitalizace. Po extrakci metadat může následovat rozhodnutí, zda digitální objekt odpovídá našim požadavkům na formát nebo jiným. Pokud ano, je objekt uložen, pokud ne, lze provést např. migraci nebo jinou úpravu. Kompletní postup zpracování digitálních objektů na vstupu do LTP systému vypadá většinou, jak ukazuje Obrázek 3.



Obrázek 3 – Nejčastější podoba workflow použití nástrojů na dlouhodobou ochranu.

[84] http://wiki.opf-labs.org/display/TR/Home
[85] http://www.digitalpreservation.gov/tools/

JHOVE1/JHOVE2[86] (*JSTOR/Harvard Object Validation Environment*) – je velmi úspěšný a všude využívaný nástroj, vyvinutý na Harvardské univerzitě ve spolupráci s organizací JSTOR. Cílem bylo automatizovat identifikaci, validaci a charakterizaci digitálních objektů. Nástroj pracuje s několika datovými formáty (AIFF, ASCII, BYTESTREAM, GIF, HTML, JPEG2000, JPEG, PDF, TIFF, UTF8, XML, ZIP a WAV). Je v mnoha ohledech nastavitelný, např. jaká je podoba výstupu (délka, výstupní formát txt nebo XML, obsah záznamu), způsobu jeho práce s objekty. Problémem je, že JHOVE nepodporuje běžné formáty kancelářských dokumentů, kromě PDF. To může působit problémy hlavně v archivech, které tyto dokumenty budou od institucí dostávat.

DROID[87] *(Digital Record Object Identification)* – je nástroj vyvinutý Národním archivem Velké Británie. Provádí automatickou identifikaci formátů jak u jednotlivých objektů, tak hromadně. Provádí také některé úkony charakterizace, dokáže poskytnout informaci jak interní struktura konkrétního objektu odpovídá standardu, ke kterému se hlásí. Výstupem je informace o konkrétní verzi formátu digitálního objektu. K tvorbě výstupu využívá databázi registru formátů PRONOM, ke které je připojen. PRONOM obsahuje informace o formátech, které DROID identifikuje. Nutno podotknout, že ač JHOVE i DROID/PRONOM používají k identifikaci stejný postup (dle přípon, podpisů a sekvencí bytů), výsledky jsou často různé. Jde o známé problémy a je s nimi při práci v LTP systémech počítáno.[88] DROID často využívá pouze rozlišení dle přípony, což není dostatečné. Mnoho různých formátů může mít stejnou příponu, včetně variant stejného formátu. Ve své nové verzi 6 umí DROID také kontrolovat integritu jednotlivých digitálních objektů i celých složek. V případě změny hlásí problém.

AONS[89] (*Automatic Obsolescence Notification System*) – je systém k automatickému upozornění na zastarávání formátů v repozitářích. Dokáže upozornit správce repozitáře na hrozící nebezpečí. Vyvinut byl Národní knihovnou Austrálie ve spolupráci s organizací APSR (*Australian Partnership for Sustainable Repositories*). Nástroj je propojen s registrem formátů PRONOM. Podobný automatizovaný systém na sledování prostředí vně repozitáře (komunita, technologie, proměny formátů a systémů apod.) se vyvíjí v evropském projektu SCAPE[90] v pracovní skupině *Automated Watch*. Bude navazovat na nástroj plánování ochrany PLATO.

NZME[91] (*New Zealand Metadata Extractor*) – jde o jeden z prvních nástrojů pro dlouhodobou ochranu vůbec. Vytvořen byl v Národní knihovně Nového Zélandu v roce 2003 se záměrem mít nástroj, který dokáže z digitálních objektů vyextrahovat ochranná metadata (převážně technická). Nástroj se dodnes vyvíjí a v mnoha LTP systémech doplňuje JHOVE. Oba nástroje umí některé formáty zároveň, některé umí pouze NZME (formáty sady MS Office).

FITS[92](*The File Information Tool Set*) – jde o nástroj, který provádí identifikaci, validaci a také extrakci metadat z různých typů digitálních objektů. FITS spojuje různé obecně známé nástroje

[86] http://hul.harvard.edu/jhove/
[87] http://droid.sourceforge.net/
[88] Možnost nastavit pravidla obcházení určitých chyb, které nástroje hlásí a nejsou přitom skutečné.
[89] http://www.apsr.edu.au/aons2/
[90] http://www.scape-project.eu/
[91] http://meta-extractor.sourceforge.net/
[92] http://code.google.com/p/fits/

(JHOVE, ExifTool, NZME, DROID, FFident aj.) do jednoho, normalizuje a upravuje jejich výstupy a také hlásí jejich chyby. Vytvořen byl na univerzitě v Harvardu pro použití v jejich LTP systému.

FIDO[93] – nový (2010) velmi rychlý nástroj na identifikaci formátů digitálních objektů. Používá podpisy z databáze PRONOM.

XCEL a XCDL[94] – v projektu PLANETS vytvořené XML jazyky pro popis charakteristických vlastností digitálních objektů. XCDL (*eXtensible Characterisation Description Language*) slouží pro popis vlastností. XCEL (*eXtensible Characterisation Extraction Language*) slouží pro extrahování vlastností z digitálních objektů. Jde o jeden z nástrojů, který je schopen automatickou cestou z objektu získat jeho vlastnosti a zapsat je v XML podobě, která se dá dále využívat např. pro validace nebo vytváření technických metadat. Oba jazyky tedy umožňují vytvořit popis vlastností objektů a tyto pak jednoduše automaticky porovnávat, např. před migrací a po migraci.

XENA[95] – je nástroj na určení datových formátů a na převod digitálních objektů do formátů určených k dlouhodobé ochraně. Je součástí australského balíku nástrojů pro podporu logické dlouhodobé ochrany DSPS (*Digital Preservation Software Platform*). Vyvinuto v národním archivu Austrálie.

3.2.5.2 Registry formátů

S validací formátů jsou velmi blízce spojeny registry formátů. Nástroje na validaci na tyto online registry odkazují a díky tomu mají k dispozici nejaktuálnější technické informace o konkrétním formátu. Registry obsahují údaje o formátech, jejich dokumentaci, vlastnostech, nástrojích na migrace, HW a SW prostředí pro jejich zobrazení. Bohužel ne vždy jsou tyto údaje k dispozici. Registry by měly odpovědět mj. na následující otázky (odpovídá registru PRONOM):

- Mám digitální objekt, co je to za formát?
- Digitální objekt uvádí, že jde o formát X, je to opravdu formát X?
- Mám objekt ve formátu X a chci jej převést na formát Y, jak?
- Mám digitální objekt ve formátu X, jaké má vlastnosti?
- Mám digitální objekt ve formátu X, jaká k němu existuje dokumentace?
- Mám digitální objekt ve formátu X, jaké je s ním spojeno riziko?
- Mám digitální objekt ve formátu X, jak a čím jej mohu zobrazit?

Již v roce 2005 Tobias Steinke [STEINKE, 2005, s. 5] popsal potřebu mezinárodní databáze, která by obsahovala informace o jednotlivých datových formátech – tedy registr formátů. Každý formát by měl identifikátor, doprovodné informace o nástrojích, závislostech apod. Pomocí identifikátoru by bylo možné odkazovat na konkrétní formát z ochranných metadat. V roce 2005 byla již v provozu databáze PRONOM v Národním archivu Velké Británie, která později tuto vizi naplnila a je dnes využívána všemi LTP systémy.

[93] http://www.openplanetsfoundation.org/blogs/2010-11-03-fido-%E2%80%93-high-performance-format-identifier-digital-objects
[94] http://planetarium.hki.uni-koeln.de/planets_cms/about-xcl
[95] http://dpsp.sourceforge.net/

Před vznikem PRONOMu měla např. národní knihovna Nizozemí v rámci LTP systému DIAS svůj registr formátů, podobně jako novozélandská národní knihovna. Tento přístup je dnes v LTP systémech (Rosetta a SDB) běžný. Tzv. lokální registr formátů (nebo formátová knihovna) je oproti globálním registrům doplněn o informace lokálního charakteru. Může jít např. o informace vypršení licencí k jednotlivým SW aplikacím, což může mít dopad na jednotlivé datové formáty uložené v LTP systému.

PRONOM[96] – registr formátů vyvinutý Národním archivem Velké Británie jako databáze znalostí o technických informacích. Obsahuje informace o zhruba 600 formátech souborů a 250 SW nástrojích. V podstatě vždy se používá v součinnosti s nástrojem na identifikaci formátů DROID. PRONOM se využívá ve všech LTP systémech jak komerčních, tak open source i přesto, že má spoustu nedostatků, které jsou všeobecně známé. Jedním z nich je nedostatečný popis formátů. Jednotlivé záznamy formátů mají místa na doplnění údajů o dokumentaci, kódování, právech s formátem spojeným, kompresi, ta ovšem jsou plněna velmi zřídka. Naproti tomu ***GDFR/UDFR***[97] (*The Global/Unified Digital Formats Registry*), další z registrů formátů, popisuje formáty více do detailu.

Preserv2[98] – britský registr formátů vytvořený jako sémanticky bohatší varianta PRONOMu. Data v registru jsou z různých zdrojů, např. PRONOM, dbpedia aj.
V projektu PLANETS vznikl ***Planets Core Registry***, který obsahuje popisy jednotlivých datových formátů a také SW nástrojů, včetně informace o jejich vhodnosti na dlouhodobou ochranu jednotlivých typů datových formátů. Nástroje je tak možné porovnat mezi sebou a rozhodnout se o jejich využití pro konkrétní případ. Core Registry rozšiřuje registr PRONOM. Velmi dobré je, že popisy v PLANETS registru jsou doplňovány z výsledků různých testu v online prostředí PLANETS Testbed [BONIN, 2009, s. 7].

3.2.5.3 Nástroje na plánování dlouhodobé ochrany digitálních dat

Nejznámější volně dostupné nástroje na plánování dlouhodobé ochrany jsou výstupy projektu PLANETS, konkrétně jde o PLANETS testbed a PLATO. Jsou to aplikace na testování migrací, vytváření plánů ochrany, jejichž výstupy lze zakomponovat do workflow LTP systému. Obě aplikace mají svůj prapůvod v evropském projektu DELOS[99] (2002-2006), v jehož rámci existovala i sekce pro dlouhodobou ochranu digitálních dat. V této sekci vznikly první návrhy a prototypy obou nástrojů. Jejich vývoj poté pokračoval v projektu PLANETS. Dalším typem nástrojů jsou vlastní LTP systémy, které obsahují modul plánování. Z volně dostupných můžeme uvést nástroj Archivematica nebo RODA – více viz kapitola 6.3.3.

PLATO[100] – nástroj na plánování dlouhodobé ochrany, který umožní organizaci vytvářet tzv. plány ochrany, které lze automaticky provádět ve spojení s LTP systémy nebo se systémy uložení dat. Jde o online i offline prostředí, kde lze definovat požadavky na ochranu – způsoby, metody,

[96] http://www.nationalarchives.gov.uk/PRONOM/BasicSearch/proBasicSearch.aspx?status=new
[97] UDFR je následovníkem GDFR, který má propojit PRONOM a GDFR. Funguje od začátku března 2012.
[98] http://p2-registry.ecs.soton.ac.uk/
[99] http://www.delos.info/
[100] http://www.ifs.tuwien.ac.at/dp/plato/intro.html

zachování nebo pominutí signifikantních vlastností různých typů dat. Vstupem do nástroje může být manuální nastavení nebo upload předem vytvořené myšlenkové mapy, kterou si systém PLATO přeloží do jednotlivých požadavků. Nástroj také podporuje porovnání testovacích výstupů jednotlivých metod ochrany nebo výstupů z různých nástrojů, které podporuje (ImageMagick aj.). Na základě plánu ochrany je možno definovat tzv. ochranné aktivity a v systému je provádět. Např. *digital asset management* systém EPrints je schopen od roku 2009 tyto plány ochrany pro jednotlivé typy dat (např. JPEG, GIF aj.) přímo provádět. EPrints je tak díky PLANETS doplněn o modul plánování ochrany a odpovídá více OAIS referenčnímu rámci. Stává se z něj jednoduchý LTP systém. Nástroj PLATO byl také rozsáhle testován v NK ČR v letech 2010 a 2011 v rámci institucionálního výzkumného záměru NK ČR jako výzkumné instituce. Testy pro jednotlivé datové formáty (JPEG, JPEG2000, PDF, TXT a DOC) prováděli pracovníci Odboru digitální ochrany za vedení Andrey Fojtů.

PLANETS Testbed[101] – webová aplikace, která poskytuje prostředí na vědecké experimenty v oblasti dlouhodobé ochrany digitálních dat. Přes webový prohlížeč nabízí a kombinuje data, SW, HW nástroje k testování různých metod dlouhodobé ochrany pro různé typy digitálních objektů. Výsledky lze podrobně manuálně i automaticky porovnávat. Jedním z cílů PLANETS Testbed je vytvářet kontinuálně, na základě experimentů, sdílenou znalostní bázi výkonů a možností různých nástrojů použitelných na procesy digital preservation. Uživatel má zároveň přístup k výsledkům experimentů, může tak zjistit konkrétní informace o migraci konkrétních formátů dat s konkrétním nástrojem i bez experimentu.

CRiB [102]– jde o nástroj vyvíjený mezi lety 2006-2008 v Portugalsku pro použití v národním archivu. Je do jisté míry obdobou nástroje PLATO a umožňuje porovnávat migrace, posuzovat jejich výsledky a tak získávat návrhy na nejlepší postup pro jednotlivé formáty. Nástroj lze použít i na plánování ochrany a extrakci metadat. Vývoj od roku 2011 opět pokračuje spolu s open source systémem RODA, se kterým tvoří funkční celek od počátku.

3.3 Referenční rámec OAIS – obecný popis

Lze říci, že obecný vývoj využívání, tvorby a uložení digitálních dat, začínal budováním digitálních knihoven pro zpřístupnění. Všeobecně přijímaný popis architektury digitálních knihoven nazývaný „*Kahn-Wilenski*“, podle autorů známého článku z roku 1995 [KAHN a WILENSKI, 1995], obsahoval jako jednu ze svých částí repozitář, tj. prostor, kde jsou digitální data ukládána pro další využití (pro přístup k datům). Tento repozitář ovšem sloužil jen pro statické uložení dat a celkový model architektury neřešil cíleně problematiku dlouhodobé ochrany digitálních dat. Model se zaměřoval více na strukturu a na operace v repozitáři prováděných na digitálních objektech, než na specifikaci nároků a vlastností digitálních repozitářů samotných. Bylo proto potřeba rozpracovat tuto část modelu architektury digitálních knihoven, která se z dnešního pohledu zdá býti právě tou nejdůležitější a klíčovou, tedy repozitář. Dlouhodobá ochrana dat v repozitáři umožní zpřístupnění obsahu repozitáře pomocí aplikací zpřístupnění v blízké nebo vzdálenější budoucnosti.
Na konci 90. let 20. století se proto stále více zvětšoval zájem o vytvoření standardního rámce, který by popisoval funkční komponenty a procesy vlastní většině archivů určených pro

[101] http://testbed.planets-project.eu/testbed/
[102] http://redmine.keep.pt/

dlouhodobé ukládání dat v digitální formě. Bylo nutné vytvořit obecný rámec, který by byl šířeji použitelný a stanovil společná východiska a jazyk pro další vývoj archivních systémů. Přibližně takto znělo zadání, které vzniklo na základě potřeby vývoje datového standardu pro podporu vesmírného výzkumu. Bylo zadáno jako potřeba *Consultative Committee for Space Data Systems* (CCSDS), což je organizace pro mezinárodní spolupráci vesmírných agentur. Na začátku této snahy navrhlo CCSDS referenční model, který měl ukotvit terminologii a koncepty pro popis a porovnání datových modelů a archivních architektur; popsat podstatné entity a vztahy mezi nimi v archivním prostředí; vysvětlit klíčové funkce a informační komponenty archivního systému. Práce CCSDS vyústila v květnu 1999 ve vydání referenčního modelu OAIS. Na Internetu jsou dostupné i různé pracovní verze z roku 1998[103]. Referenční model OAIS je koncepční rámec pro obecný archivní systém věnovaný problému ochrany a správy přístupu k digitálním informacím v dlouhodobém měřítku. Upravená verze modelu byla publikována v roce 2002 – viz [CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEMS, 2002]. Na základě verze z roku 2002 byl referenční rámec v roce 2003 vyhlášen za mezinárodní normu ISO 14721:2003[104]. Další menší úprava textu proběhla v roce 2007. Poslední, tentokrát již více přepracovaná a doplněná verze ISO, je z roku 2009.

Referenční model OAIS se ukázal jako velmi životaschopný a je dnes široce implementován a využíván v paměťových institucích, tedy převážně mimo původní komunitu vesmírného výzkumu. Je ideální svou obecností a tím tedy i možností implementace. Velmi podstatné je, že OAIS referenční rámec obsahuje a definuje terminologický slovník pro oblast dlouhodobé ochrany a archivů/repozitářů, který je srozumitelný široké odborné veřejnosti. Dále definuje informační model a také funkční model digitálního archivu. Tedy popisuje klíčové procesy probíhající v digitálním archivu a jeho funkční komponenty. Informační model OAIS popisuje metadata potřebná ve všech komponentách archivního systému, a na všech stupních procesu archivace pro dlouhodobé uložení digitálního objektu. Referenční rámec OAIS tak slouží jako *„high-level metadata framework for digital preservation.“* [OCLC/RLG WORKING GROUP ON PRESERVATION METADATA, 2002, s. 10] Tento metadatový rámec je nezávislý na typu digitálního objektu, na technologii dlouhodobé ochrany i samotného archivu. Díky tomu se dnes od obecného rámce daného OAIS odvíjejí veškeré koncepty metadat pro dlouhodobou ochranu, tzv. ochranných metadat, např. PREMIS – viz kapitola 4.
Model jako takový je použitelný pro všechny datové archivy a organizace nakládající s informacemi určenými pro dlouhodobou ochranu. OAIS také neřeší procesy mimo vlastní archiv, neposkytuje návod na tvorbu digitálních dat apod. To popisuje např. DCC *Curation Lifecycle Model*, který se zabývá celým životním cyklem digitálního objektu – více popsán např. v [HARVEY, 2010, s. 34-37].

Referenční model OAIS je primárně zaměřen na digitální informace, a to na primární formu archivované informace (digitální objekt) a na informace podpůrné pro digitální nebo fyzické archivované materiály. Nicméně model lze aplikovat i na nedigitální informace (např. fyzický vzorek), ovšem ochrana a architektura takových informací není rozebrána do detailu. Referenční model OAIS mj. [CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEMS, 2002, s. 1-1]:

[103] Viz např. http://nssdcftp.gsfc.nasa.gov/standards/nost/isoas/us12/CCSDS-650.0-W-3.pdf.
[104] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=24683

- poskytuje rámec pro pochopení a zvýšení povědomí o archivních konceptech a standardech potřebných pro dlouhodobou ochranu digitálních informací a přístup k nim;
- poskytuje koncept potřebný pro nearchivní instituce k tomu, aby byly užitečnými účastníky v procesech ochrany;
- poskytuje rámec, včetně terminologie a konceptů, pro popis a porovnání architektury a procesů již existujících i budoucích archivů;
- poskytuje rámec pro popis a porovnání různých strategií a technik dlouhodobé ochrany;
- poskytuje základ pro porovnání datových modelů digitálních informací ochraňovaných v archivech a také pro diskuzi o tom, jak se mohou datové modely a zásadní informace během doby měnit;
- poskytuje základ pro vývoj relevantních standardů nebo specifikací pro podporu dlouhodobé ochrany digitálních dat.

3.3.1 Referenční model OAIS a OAIS digitální archiv

Referenční model OAIS se zabývá celou škálou ochranných funkcí na různých úrovních, jako jsou vkládání objektů do archivu, archivní uložení, data management, přístup a šíření informací. Zabývá se rovněž migrací digitálních informací na nová média a formy, data modelem použitým na reprezentaci informací, rolí SW v ochraně informací a výměnou digitálních informací mezi archivy. Určuje zároveň jak interní, tak externí pracovní prostředí digitálního archivu.

Hlavní koncept referenčního modelu spočívá v OAIS (*Open Archival Information System*). Výraz *open* odkazuje na skutečnost, že referenční model byl vytvořen a volně publikován na veřejných fórech, na kterých mohl kdokoliv zainteresovaný participovat.[105] *Archival information system* je: *„… archiv sestávající z organizace lidí a systémů, kteří přijali odpovědnost za ochranu informací a jejich zpřístupnění určité komunitě. To představuje určité odpovědnosti, definované v rámci OAIS, a které dovolují takový OAIS archiv odlišovat od ostatních systémů popisovaných slovem Archiv.“* [CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEMS, 2002 s. 1-11]

Z výše uvedeného vyplývají dvě primární funkce archivních repozitářů: za prvé ochránit informaci a zajistit její dlouhodobou odolnost; za druhé poskytnout k archivované informaci přístup v závislosti na potřebách uživatelů nebo komunity. Aby tedy repozitář odpovídal OAIS, měl by obzvláště [LAVOIE, 2004, s. 3]:

- vyjednávat a získávat informace od producentů informací;
- zajistit dostatečnou kontrolu informací, aby se plnily podmínky dlouhodobé ochrany;
- vymezit hlediska využití uživatelskou komunitou;
- zajistit, aby byly chráněné informace nezávisle srozumitelné uživatelům v tom smyslu, že je možné informaci rozumět a vnímat bez využití asistence producenta této informace;
- dodržovat strategii a procedury a tím zajistit to, že informace je ochráněna proti všem nepředvídatelnostem;
- zpřístupnit chráněné informace uživatelské komunitě.

Tyto odpovědnosti bere za vlastní valná většina digitálních repozitářů, i když ne zcela. Znamená to, že jsou s OAIS kompatibilní, neznamená to ale automaticky, že OAIS zcela odpovídají a provádějí všechny procesy tak, jak je OAIS definuje – více viz kapitola 6.3.5. Tak tomu většinou není. OAIS plně odpovídají pouze LTP systémy, které mají modul plánování ochrany.

[105] Výraz neznamená, že přístup do archivu OAIS je neomezený.

Opravdová síla OAIS referenčního rámce spočívá ve specifikování procesů probíhajících v jednotlivých modulech a mezi nimi. Jde vlastně o seznam funkčních procesů pro implementaci konkrétního digitálního repozitáře, aniž by byla specifikována jeho architektura. OAIS nerozlišuje mezi repozitářem (archivem) národní knihovny, univerzitní knihovny, nebo archivem státní instituce.

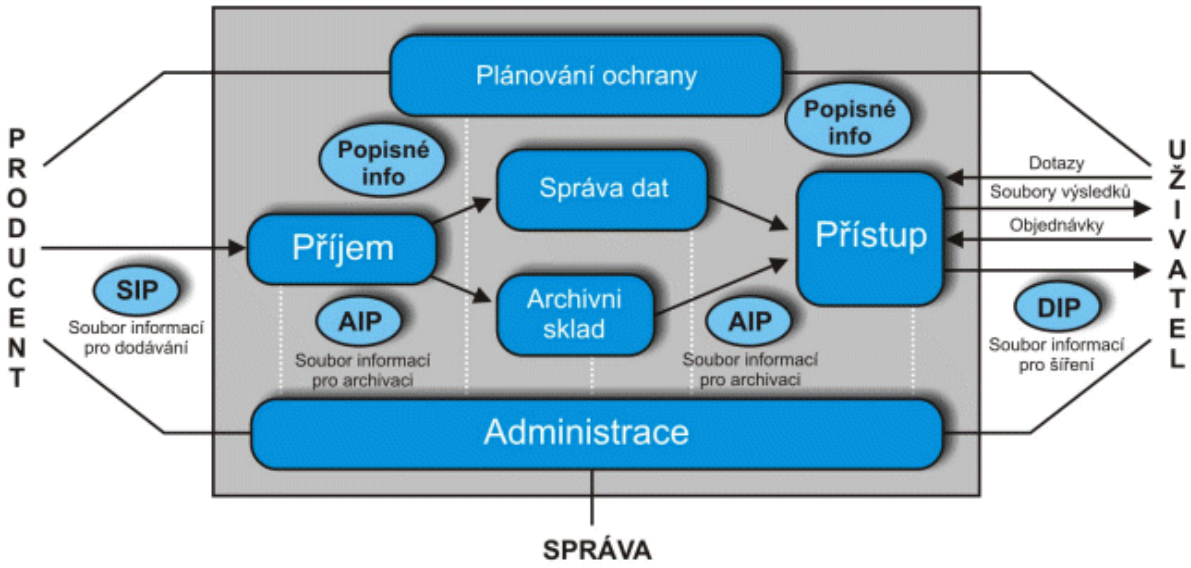
3.3.2 OAIS balíčky a moduly

OAIS archiv pracuje v prostředí formovaném vztahy mezi čtyřmi entitami: producenty (*producers*), uživateli (*consumers*), správci (*managers*) a archivem samotným. V archivu jsou digitální objekty v podobě tzv. informačních balíčků (viz níže). Producenti dat poskytují informace (data a metadata), která jsou v archivu uložena. Uživatelé tyto uložené informace (data a/nebo metadata) využívají. Mezi uživateli hraje významnou roli cílová uživatelská komunita (*designated community*), pro kterou především jsou archivované informace určeny a měla by je v budoucnu umět bez problémů najít, zobrazit a porozumět jim i jejich kontextu. Cílová uživatelská komunita může být považována za typ uživatele. Správci jsou zodpovědní za vznik a fungování strategií pro správu archivu, jako jsou např. výběr relevantních dat, financování a další věci související s archivem jako celkem [OCLC/RLG WORKING GROUP ON PRESERVATION METADATA, 2002, s. 9].

Klíčové funkce, jak je definuje model OAIS, jsou prováděny v šesti hlavních modulech referenčního rámce a tedy přeneseně i archivního repozitáře. Tyto hlavní funkční komponenty ukazuje Obrázek 4. Jsou to:

- ***modul Příjem (Ingest)*** – proces akceptování informačního balíčku od jeho producenta; modul je zodpovědný za přijetí dat a jejich přípravu na uložení a správu, tj. vytvoření AIP balíčku a doplnění potřebných metadat;
- ***modul Archivní sklad (Archival Storage)*** – zajišťuje, že data i jejich kontext budou uložena bezpečně; modul zařizuje uložení, správu a vyhledávání AIP balíčků;
- ***modul Správa dat (Data Management)*** – podporuje přístup k datům a jejich úpravy/správu; koordinuje popisné informace k datům spolu se systémovými informacemi používanými na podporu archivních funkcí;
- ***modul Administrace (Administration)*** – slouží ke správě procesů, funkcí a k nastavení repozitáře i jeho uživatelů;
- ***modul Zpřístupnění (Access)*** – poskytuje rozhranní mezi archivem a uživatelem (obecný uživatel nebo administrátor repozitáře); pomáhá uživateli vyhledat, získat a zobrazit požadovaný archivovaný obsah repozitáře[106];
- ***modul Plánování ochrany (Preservation Planning)*** – slouží k vytváření plánů ochrany, testování nástrojů a metod dlouhodobé ochrany a k provádění ochranných akcí.

[106] Zobrazení je již mimo archiv OAIS, hlavní funkcionalita modulu Přístup je dodat DIP balíček uživateli.



Obrázek 4 – Referenční rámec OAIS. Dle předlohy [CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEMS, 2002, s. 4-1] překreslil Martin Zhouf.

OAIS a repozitář jemu odpovídající pracují s konceptem *informačních balíčků* – viz Obrázek 4. Ty obsahují jak data určená k archivaci, tak metadata o nich a také metadata k balíčku včetně informace o zabalení *informačního balíčku*. Tedy balíček neobsahuje pouze data, ale také informace potřebné k ochraně digitálních objektů, které jsou jeho součástí – více viz kapitola 4.9.1. OAIS rozlišuje tři typy balíčků (SIP, AIP a DIP), které se od sebe samozřejmě liší a mají konkrétní úkoly v rámci repozitáře.

Submission Information Package – SIP

- je předmětem vyjednávání mezi producentem (tvůrcem informací) a OAIS archivem;
- je zasílán producentem (dodavatelem) do OAIS archivu;
- obsahuje data určená k archivaci a k nim náležející popisná, technická případně jiná metadata.

Archival Information Package – AIP

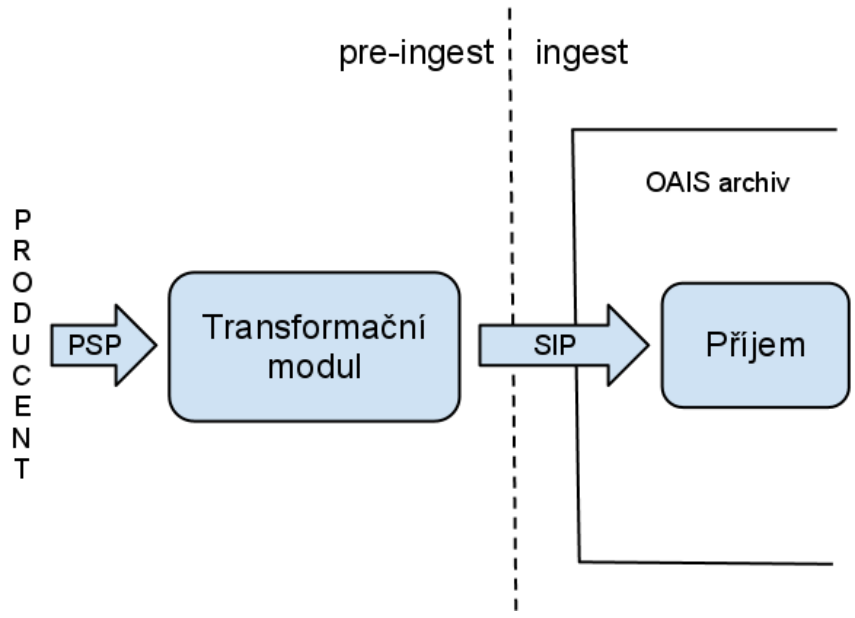
- vzniká v OAIS archivu z balíčku SIP, nejčastěji přidáním dalších nutných metadat nebo změnou struktury SIP;
- AIP je balíček informací používaný pro archivní uložení;
- obsahuje kompletní sadu tzv. *Popisné ochranné informace* (*Preservation Description Information – PDI*) pro obsahové informace.

Dissemination Information Package – DIP

- obsahuje část nebo všechny části jednoho nebo více AIP;
- vzniká na vyžádání dat z archivu uživatelem (osoba nebo aplikace);
- OAIS je rozšiřuje uživatelům (je to vlastně to, co se dostane z repozitáře ven).

Mimo specifikaci OAIS existuje ještě koncept tzv. **PSP balíčku (*producer submission package*)**. Poprvé se začal používat tento výraz na přelomu let 2008/2009 v pracovní skupině LTP, kterou

iniciovaly národní knihovny Německa a Nizozemí. Cílem skupiny odborníků z několika evropských národních knihoven bylo specifikovat obecnou funkcionalitu LTP systému tak, aby ji bylo možné použít pro výběrová řízení na tento typ systému plánovaná v různých zemích Evropy. Pracovní skupiny se účastnila i NK ČR (J. Hutař, M. Melichar). Členové skupiny rozšířili OAIS referenční rámec o další návazné části a vrstvy. Jednou z nich byl právě PSP balíček, což jsou data, která přicházejí od vydavatele. Až z těchto dat se vytváří SIP balíček pro vložení do LTP systému. To je podstatné rozšíření OAIS v oblasti příjmu (*ingestu*). OAIS popisuje pouze hotový SIP, neřeší otázku jeho vzniku. SIP dle OAIS musí být připraven na příjem (*ingest*) do archivu, což v reálném prostředí LTP systému není vždy možné. SIP totiž musí být ve struktuře a formátu, kterou LTP systém požaduje. Do této struktury je nutno jej převést, pokud ji nemá od producenta informace. PSP balíček od producenta je často v různých formátech a s různou strukturou. Obvykle není možné vydavatelům nařizovat, jaké formáty dat mají posílat. Tato změna, vlastně změna PSP na SIP, se odehraje před příjmem do archivu (*ingestem*) a probíhá v modulu/aplikaci mimo OAIS. Proto bylo nutné přijít s konceptem PSP. Množina procesů, které změnu PSP na SIP doprovázejí, se nazývá *pre-ingest* a je s ním běžně operováno v prostředí LTP systémů. Jak celý koncept vypadá v souvislosti s OAIS archivem ukazuje Obrázek 5.



Obrázek 5 – PSP balíček od producenta a návaznost na OAIS archiv.

4. Úvod do metadat

Jak vyplynulo z předchozích kapitol, ochraňovat digitální objekty znamená dobře znát jejich typické vlastnosti, které se musí zachovat dlouhodobě, aby byl digitální objekt použitelný, smysluplný, autentický a seriozní. Tyto vlastnosti částečně závisí na formátu objektu, ale velká část z nich je určena kontextem, ve kterém byly objekty vytvořeny a budou využívány [STOKLASOVÁ a HUTAŘ, 2007, s. 94]. Informace o vlastnostech, kontextu jsou uloženy právě v metadatech.

Je zdokumentováno, že pojem *metadata* použil vůbec poprvé Jack E. Myers v roce 1969 [GREENBERG, 2005, s. 19].[107] Myersův termín se ve smyslu „data o datech" rozšířil do komunit IT vývojářů, statistiků, databázových odborníků, a později i mezi knihovnickou obec. Ve větší míře se pojem metadata objevuje v polovině 80. let 20. století, kdy začaly vznikat elektronické archivy digitalizovaných textů. Vedle knihovníků používali pojem metadata také velmi často komunity zapojené ve správě a sdílení dat vzniklých v geografii. Pro tyto komunity pojem označoval soubory oborových nebo průmyslových standardů stejně jako doprovodné informace k vlastním datům uloženým v informačních systémech [HOWARTH, 2005, s. 40]. S příchodem Internetu se začalo mluvit o metadatech pro online zdroje a termín metadata se v komunitě paměťových institucí stal běžně používaným pro jakékoliv dodatečné, strukturované informace, které se pojí k digitálním, ale i analogovým dokumentům.

Obecně je metadaty myšlena jakákoliv doprovodná informace k nějakému objektu, která jej popisuje, usazuje do kontextu, umožňuje jeho vyhledávání, zpřístupnění. Metadata se používají na popis *informačních objektů*. Za informační objekt můžeme považovat cokoliv, k čemu lze přistupovat a s čím lze zacházet jako s diskrétní entitou. Takovým objektem může být jednotlivina, nebo soubor jednotlivin (obrazy, HTML stránky, videa, muzeální objekty, sbírky, služby, události, osoby, místa nebo i samotné metadatové soubory). Pokud se jedná o informační objekt v digitální podobě, jde vždy o sekvenci bitů (bitstream). Bitstream sám o sobě nelze považovat za samopopisný. Pokud nevíme, co konkrétní bitstream reprezentuje, neznáme kontext, velmi těžko můžeme takový bitstream „rozluštit" a použít v budoucnu. Může být jednoduché bitstream rozkódovat a zobrazit, aniž bychom znali kontext a doplňující informace, např. v případě jednoduchého bitstreamu jednoduchého lineárního textu [ROTHENBERG, 1999, s. 8]. Ve většině případů to ale možné není, pokud nemáme odpovídající metadata. Tady hrají metadata největší roli, přidání kontextových informací (popisných, technických aj.) k bitstreamu.
Ať má informační objekt fyzickou nebo intelektuální formu, lze vždy popsat tři jeho vlastnosti: obsah, strukturu a kontext. Tato tři hlediska mohou být a jsou popisována metadaty [BACA, 2008, s. 2]. V prostředí, kde mají uživatelé bezprostřední přístup k informačním objektům, metadata [BACA, 2008, s. 6]:

- potvrzují autentičnost a úroveň kompletnosti obsahu;
- stanovují a dokumentují kontext obsahu;

[107] Myers použil termín „metadata" pro popis svých produktů, které měly spojitost s jeho MetaModelem a se společností, kterou pod názvem Metadata založil. Na slovo „metadata" v určitém kontextu také dodnes vlastní ochrannou známku.

- identifikují a využívají strukturální vztahy existující uvnitř, anebo mezi informačními objekty;
- poskytují možnost zpřístupnění pro různé typy uživatelů.

Již od devadesátých let byly knihovny na celém světě, i v ČR, tahouny v oblasti digitalizace. Jedním z důvodů byla připravenost, která spočívala v provedené automatizaci, existenci standardů na popis fyzických dokumentů (standard MARC) a celková vstřícnost knihovníků k technickým řešením. Celé období od počátku do poloviny prvního desetiletí 21. století je typické velkým důrazem na digitalizaci a relativně velkými objemy zdigitalizovaných dat, která vznikla.[108] K této množině dat vzniklých digitalizací se díky rozvoji technologií a Internetu přidávaly velmi rychle tzv. *digital-born* dokumenty.

Digitalizace samotná je jen prvním předpokladem k odpovídajícímu využití digitálních zdrojů. Miliony zdigitalizovaných obrazů bez základního popisu a indexace těchto popisných údajů, jsou pro běžného uživatele v podstatě nedostupné a nepoužitelné. Druhým předpokladem pro využití jak jej známe běžně z aplikací zpřístupnění (digitálních knihoven) dostupných na Internetu, jsou tedy metadata. Vzhledem k proměně konceptu práce s dokumentem, od tradičního k elektronickému, která zasáhla všechny knihovnické procesy, došlo i k proměně pohledu na katalogizaci a popis dokumentů samotných. Právě v popisu dokumentů jsou knihovny tradičně silné a lze jej považovat za hlavní výhodu oproti objevujícím se soukromým aktivitám na poli zpřístupnění a zprostředkování informací. Pokud se knihovny budou soustředit na popis a odpovídající zpřístupnění na základě metadat, lze očekávat, že knihovny nepostihne neblahý osud, který jim mnozí prorokují.

Popisná metadata jsou nezbytná pro využití digitálních zdrojů i jejich provázání s různými nadstavbami a službami, které digitální knihovny samotné „překračují“. Další oblastí kde metadata hrají nezastupitelnou roli, je oblast správy digitálních dokumentů a oblast dlouhodobé ochrany digitálních dat. Paměťové instituce se velmi dlouho soustředily na popis digitálních dokumentů, nezabývaly se ovšem hledisky jejich ochrany. Právě v této oblasti jsou metadata životně důležitá, protože mohou pomoci uchovat jakoukoliv informaci, která se k elektronickému dokumentu pojí (informace technické, procesně/administrativní, informace o právech a omezeních i o struktuře dokumentů). Tato klíčová vlastnost metadat, schopnost uchovat technické a jiné informace, se začala brát do úvahy relativně pozdě, až když se objevily první problémy a následné úvahy o tom, co v budoucnosti bude s miliony nově zdigitalizovaných dokumentů. Vedle popisných metadat pro popis a vyhledávání tak existují další typy metadat. Můžeme mluvit o metadatech administrativních, ochranných, technických, vzdělávacích, metadatech práv, webových metadatech v HTML kódu aj. – viz kapitola 4.3.

4.1 Definice metadat

Definice organizace NISO popisuje metadata jako: „*… strukturovanou informaci, která popisuje, vysvětluje, lokalizuje nebo jinak ulehčuje získání, použití nebo správu informačního zdroje*.“ [NISO, 2004, s. 1]

V podstatě totožná je definice pracovní skupiny CC:DA (*Comittee on Cataloging: Description and Access*), která existovala koncem 90. let 20. století při ALCTS (*The Association for Library*

[108] V porovnání s dnešní situací, kdy ve většině zemí probíhá nebo se rozjíždí tzv. masová digitalizace, jsou ovšem objemy dat vytvořené digitalizací do roku 2005 malé.

Collections & Technical Services). Metadata dle této skupiny jsou: „*strukturovaná, kódovaná data popisující vlastnosti entit nesoucích informace tak, aby pomohla s identifikací, nalezením, posouzením a správou popisovaných entit.*“ [ASSOCIATION FOR LIBRARY COLLECTIONS & TECHNICAL SERVICES, 2000]
Dublin Core Metadata Initiative popisuje ve svém slovníku pojmů z roku 2005 [WOODLEY, 2005] metadata jednoduše jako „*strukturovaná data o datech*“.
Rozšířenou definicí je také definice Tima Berners-Lee, který akcentuje skutečnost, že až Internet dal impuls většímu rozvoji a výměně metadat v té podobě, jak je známe dnes. Jeho definice míří na metadata náležející webovým zdrojům, ale nevynechává ani zdroje ostatní. „*Strojům srozumitelné informace o webových zdrojích nebo dalších věcech.*“ [BERNERS-LEE, 1997] Vedle definice Berners-Lee uvádí několik axiomů o metadatech. Jedním z nich je, že „*metadata jsou data*“. Metadata jsou opravdu data (digitální objekty), která se často ukládají s objekty, které popisují.

Je několik skutečností, na kterých se všechny definice metadat shodnou. Metadata jsou data o datech, tj. reprezentují jiná data. Podstatné je, že jde o strukturovaná data. Právě slovo „strukturovaná“ je důležité pro odlišení jakýchkoliv dodatečných informací od metadat. Ta se totiž strukturou vyznačují. Význam struktury stoupal s rozvojem možností strojového zpracování počítači. Tvůrci metadat jsou si vědomi toho, že čím více je informační objekt strukturovaný, tím lépe je možné strukturu využít k vyhledávání, správě a manipulaci s digitálními objekty. Z definic je také patrné, že metadata mají určitý cíl, umožnit vyhledávání, případně správu objektů.

4.2 Metadatové schéma

Základním pojmem v oblasti metadat je metadatové schéma. Obecné definice popisují metadatové schéma jako: „*sadu metadatových elementů určenou ke konkrétním účelům, jako např. k popisu konkrétního typu informačního zdroje.*“ [NISO, 2004, s. 2] Podobně Priscilla Caplan, která metadatové schéma definuje jako: „*sada elementů metadat a pravidel pro jejich použití, která byla definována za konkrétním účelem.*“ [CAPLAN, 2003]
Konceptualizace metadatových schémat je mj. dána normou ISO/IEC 11179[109], která má šest částí publikovaných mezi lety 2003-2005. Pro vývoj nebo specifikaci schématu jsou nejpodstatnější části 1 (*Framework*) a 4 (*Formulation of Data Definitions*) normy. Část 1 popisuje hlavní myšlenky spojené s tvorbou elementů, jejich hodnot, část 4 poskytuje návod jak vytvořit datové definice. Příbuzné normy jsou ISO/IEC 20943 (*Information technology – Procedures for achieving metadata registry content consistency – Part 1: Data elements*)[110] a také ISO/IEC 20944[111] (*Information technology – Metadata Registry Interoperability and Bindings (MDRIB)*).

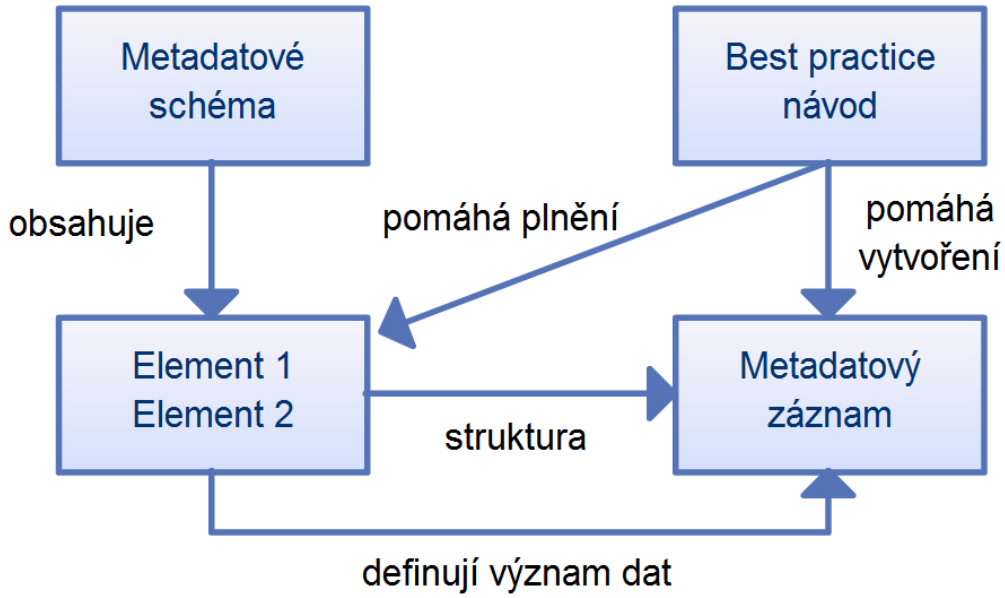
Konkrétní schéma specifikuje množinu elementů, které lze pro tvorbu metadatového záznamu využít, jejich strukturu, sémantiku, syntax záznamu a případně možnosti plnění hodnot elementů.
Syntax schématu určuje, jak jsou zapsány elementy ve strojem čitelné podobě. Vlastně tak specifikuje formát pro výměnu metadat mezi jednotlivými systémy a také podobu uložení metadat v lokálním systému. Některá schémata nemají předepsanou žádnou syntax, nebo naopak

[109] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=35343
[110] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=34343
[111] http://metadata-standards.org/20944/index.html

mají pro syntax více možností (SGML, HTML, XML[112], RDF). Hodnoty elementů, tedy vlastní obsah metadatového záznamu, mohou být tvořené volně, nebo kontrolovaně. Kontrolované plnění může být omezeno pravidly, jako jsou např. AACR2. Kontrolované hodnoty mohou být dané zcela přesně, a to kontrolovanými slovníky, klasifikacemi a identifikátory. Kontrolovaný slovník obsahuje všechny hodnoty, které může element nebo atribut nabývat. Slovník může být velmi rozsáhlý, ale také minimalistický, se dvěma položkami (označení typu dokumentu „monograph“ nebo „periodical“). Slovníky se nehodí pro všechny elementy, typicky nejsou vhodné např. pro názvy, jména. Pro tyto elementy se používají tzv. autoritní záznamy, které jsou svého druhu kontrolovaným slovníkem. Omezení hodnot pomocí slovníků a autorit výrazně napomáhá vyhledávání a tedy správě digitálních objektů podle určitých polí/elementů. Podobně je to s výše uvedenými klasifikacemi.
Identifikátory jsou speciální formou metadat, případně hodnotou elementů. Jde o řetězec znaků, který jednoznačně určuje konkrétní entitu s různými cíly. Některé identifikátory se týkají celých děl, některé pouze jejich jednotlivých částí nebo různých vyjádření [CAPLAN, 2003, s. 29]. Existují identifikátory určené pouze pro fyzické dokumenty, nebo naopak pouze pro dokumenty digitální.
Ze zástupců identifikátorů lze uvést známé ISBN, ISSN a čárový kód, které se nejčastěji (ne výlučně) používají pro fyzické dokumenty a pro bibliografický popis. Pro digitální dokumenty jsou to např. DOI, ARK, URN:NBN, PURL, UUID aj.

U schémat je podstatné mít jasný model použití. Pro které entity lze použít jaké elementy apod. Proto se schémata někdy popisují v tzv. *datovém slovníku*, který popisuje jednotlivé elementy, jejich použití, význam (sémantiku) a možné hodnoty včetně kontrolovaných slovníků – viz např. datový slovník schématu PREMIS. Ten u každého elementu uvádí, zda lze použít pro bitstream, soubor, nebo reprezentaci, příklady, význam elementu i jeho podstatu.
Různé způsoby použití a pravidel pro jedno konkrétní schéma, v závislosti na konkrétním projektu nebo typu objektu, pak definuje tzv. *aplikační profil*. Ten je vhodný zvláště tam, kde se používá pro popis jednoho dokumentu více schémat nebo pokud schéma nemá jasně daná pravidla plnění elementů, případně strukturu. Jednotlivci, kteří s takovým schématem pracují a metadata vytváří, totiž vždy budou postupovat různě. Aplikační profil jim ovšem poskytne přesná vodítka na místech, kde je schéma samotné neposkytuje. Profil je často striktnější než schéma samotné, záleží na situaci nebo konkrétním projektu. Profil může taky být platný na národní úrovni. Vztah mezi schématem, metadatovým záznamem a návody ukazuje Obrázek 6.

[112] XML je nezávislé na technologii a systémech, které jsou používány na tvorbu XML souborů i na jejich zobrazení. Je srozumitelné stroji i člověku, což je velká výhoda. XML napomáhá výměně metadat, jejich validaci. Právě v devadesátých letech lze sledovat největší snahu a o vývoj SGML/XML schémat pro různé typy metadat, a to ve všech oblastech, nejen v oblasti paměťových institucí.



Obrázek 6 – Vztah metadatového schématu a metadatového záznamu [ZENG a QIN, 2008, s. 13].

Schémata mohou být v různých podobách, nejčastější je případ, kdy všechny elementy jsou uvedeny v jednom kompletním schématu (např. MODS). Lze se setkat také s možností, kdy metadatové schéma je tvořeno více soubory/schématy. Tento přístup se používá v případech, kdy schéma je rozsáhlé a víceúrovňové a je tak lepší publikovat schéma pro celek, které pouze odkazuje schémata pro jednotlivé úrovně. Takto bylo např. řešeno schéma PREMIS v jedné z jeho předchozích verzí, stejně je řešeno schéma DDI[113] (*Data Documentation Initiative*).

Za dobu, kdy se paměťové instituce zabývají metadaty, došlo ve vytváření metadatových schémat k podstatnému posunu. Zhruba do roku 2000 si každá instituce, která chtěla vytvářet metadata pro konkrétní projekt, vytvořila vlastní schéma. Důvodem bylo přesvědčení, že pouze schéma na míru bude projektu a jeho potřebám vyhovovat nejlépe. V novém tisíciletí je nově znát silný posun ke sdílení dat skrze metadata, což vede k používání stejných schémat pro různé projekty i typy institucí. Sdílení a využívání standardů je dnes hlavní směr vývoje metadat. Znamená to dát dohromady různorodé skupiny, které si roky vytvářely svá metadatová schémata a spolupracovaly dosud minimálně. Hybnou silou hovořící pro využití standardů je možnost vyhnout se zdlouhavému vývoji a testování schémat, potřeba spolupráce, využití shodných systémů, chuť ke sdílení metadat a také podpora komunity používající konkrétní standard. Využití konkrétního standardu je vždy otázkou kompromisu, protože schéma málokdy nabízí všechny možnosti, které požadujeme. I přesto by tvorba vlastního metadatového schématu, které přesně vyhoví našim nárokům, měla být až poslední možností. Výhodou této varianty zůstává snad jen absolutní volnost tvorby pro konkrétní potřeby. Tím ovšem výhody končí. V konečném důsledku je vždy lepší vybrat si již existující a rozšířený standard, což nám umožní:

- snížit náklady na vývoj a úpravy specifikace (pomůže komunita využívající konkrétní standard);

[113] http://www.ddialliance.org/

- využívat dostupné SW aplikace a nástroje, např. na automatickou tvorbu metadat se standardním výstupem, ušetří se tím v procesu vytváření metadat;
- účastnit se mezinárodních aktivit a projektů;
- využívat dostupnou dokumentaci, případně se poučit z již existujících nasazení konkrétního metadatavého schématu;
- sdílet volně informace ve formě metadat;
- zvýšit využití a viditelnost našich digitálních sbírek – napojení do kooperačních nástrojů;
- jednoduchý přenos dat z jednoho repozitáře do druhého bez potřeby tvorby nebo změny specifikace;
- řešit problémy v rámci komunity – diskusní fóra, debaty s kolegy, kteří používají stejný standard.

Digital curation a dlouhodobá ochrana digitálních dat jsou aktivity se silným mezinárodním vývojem a komunitou. Použití standardů metadat pro data je v případě, že se instituce chce těchto aktivit účastnit, naprosto klíčová.

4.3 Typy metadat a jejich zástupci

První věc, která většinu lidí napadne ve spojitosti s metadaty je popis digitálního objektu ve smyslu bibliografického záznamu. V počátcích metadat byl popis nejpodstatnější úlohou metadat. Na popisu je následně založeno hledání, identifikace, výběr a zpřístupnění konkrétního dokumentu nebo informace. Metadata, která toto dobře umožňovala, byla z pohledu katalogizátorů pokládána za dobrá. Rozvíjející se prostředí Internetu a technologií obecně ale přinesly nové možnosti popisu, které s katalogizací a bibliografickým záznamem nemají mnoho společného. Velmi brzy bylo jasné, že metadata nejsou klíčová pouze pro popis, ale i pro další aktivity spojené s digitálními objekty/dokumenty. V obecné rovině nám metadata pomáhají (volně dle [HARGREAVES, [2010]; BACA, 2008, s. 13]:

- *najít, identifikovat, rozlišit a porozumět digitálním objektům*. Efektivita vyhledávání může být velmi posílena existencí kvalitních popisných metadat, která mohou umožnit i vyhledávání napříč různými sbírkami, pokud používají interoperabilní metadatové standardy (sdílejí zásadní pole jako autor, název apod.). Konkrétně tyto údaje obsahují popisná metadata.
- *popsat a spravovat digitální objekty*. Popisná metadata popisují základní vlastnosti digitálních objektů, případně jejich fyzických předloh. Administrativní, technická a ochranná metadata, jako např. formát, velikost souboru, kontrolní součet aj. popisují technické vlastnosti objektů. Přispívají tak spolu s popisnými metadaty ke správě objektů v repozitáři, možnostem vyhledání, filtrace dle konkrétních technických vlastností a provádění následných akcích na takto vybraných množinách digitálních objektů. Velmi podstatným pro porozumění digitálním dokumentům v budoucnu je kontext jejich vzniku, který lze vyjádřit v administrativních metadatech.
- *provádět dlouhodobou ochranu digitálních objektů*. Pokud mají digitální objekty přežít jednotlivé generace technologií v použitelné podobě, jsou potřebná popisná, administrativní, technická a ochranná metadata, která jim umožní existovat nezávisle na konkrétním úložném systému. Metadata dokumentují celý životní cyklus od okamžiku vzniku digitálního objektu, včetně všech změn jeho chování nebo podstatných vlastností. Mohou ochránit proti nechtěné změně nebo narušení integrity digitálního objektu, tím že ji dokumentují. Pomáhají autenticitě

objektu, uchováním historie událostí, které se s objektem děly (tzv. *audit trail*). Na metadatech jsou založeny opatření ochrany, např. migrace aj.

- *organizovat a vyjadřovat vztahy mezi digitálními objekty/dokumenty*. Vnitřní strukturu digitálních objektů i dokumentů vyjadřují většinou strukturální metadata (např. logická a fyzická strukturální mapa standardu METS nebo vyjádření vztahů ve standardu MODS). Strukturální metadata také mohou zaznamenat vztahy s dalšími digitálními objekty (vztahy typu „is part of", „host" a jiné).
- *používat vlastní digitální objekt*. Díky údajům v metadatech práv a licencích je možné využívat digitální objekty v souladu s právy, která se k nim vztahují. Technická metadata obsahují informace o formátu, ochranná metadata o HW a SW nutném ke zpřístupnění/použití digitálního objektu.
- *vyměňovat digitální objekty*. Možnost sdílení a výměny metadat spočívá v zajištění interoperability skrz používání stejných nebo podobných standardů metadat, výměnných protokolů, případnému mapování pomocí tzv. *mapovacích tabulek* apod. Je nutné, aby různá metadatová schémata, mezi kterými chceme provádět např. výměnu metadat, měla podobnou strukturu a sémantiku hlavních elementů.
- *validovat a kontrolovat digitální objekt*. Metadata hrají důležitou úlohu v dokumentaci všech změn a poskytují tak podklady pro kontrolu autenticity a důvěryhodnosti digitálního objektu pomocí administrativních a ochranných metadat.

Pro naplnění výše jmenovaných funkcí vznikají různé metadatové standardy. Pro popis (Dublin Core, MODS), pro administrativní metadata (PREMIS aj.), pro technická metadata (PREMIS, MIX aj.), strukturální metadata (METS). Některá schémata mají své zaměření a principy fungování popsány ve speciálních dokumentech nebo v dokumentaci. Další schémata mohou vznikat nejen na popis objektů, ale i dalších prvků, např. událostí, osob (agentů), vztahů, struktury apod.

Rozdělení typů metadat existuje několik a není možné říci, že některé rozdělení je lepší než jiné. Záleží z jakého úhlu pohledu se na množinu standardů metadat autor typologie dívá. Susan Schreibman rozlišuje ve své práci *Best practice guidelines for digital collections* [SCHREIBMAN, 2007] metadata popisná, administrativní, technická, strukturální a ochranná. Naproti tomu Murtha Baca ve svém výborném úvodu do metadat [BACA, 2008, s. 9] uvádí vedle administrativních, popisných, ochranných, technických ještě metadata pro využívání digitálního objektu (*use metadata*), podobně také [HOWARTH, 2005, s. 42]. Naopak neuvádí ve svém přehledu metadata strukturální. Za metadata pro využívání digitálního objektu považuje např. vyhledávací logy, záznamy o půjčování a také metadata práv, která jsou běžně udávána jako podmnožina administrativních metadat. *Use metadata* jsou myšlena spíše pro využití v digitálních knihovnách, ne v prostředí digitálního repozitáře, který bývá velmi často nastavený tak, že z něj žádné zpřístupnění pro běžného uživatele není možné.

Pro zajímavost uveďme typy metadat, které v roce 1996 rozlišoval ve svém článku Carl Lagoze [LAGOZE, 1996]. Popisná metadata dle něj nepokrývají všechny typy informací, které je nutné o popisovaném dokumentu/objektu zaznamenat a uložit. Vedle popisných metadat uváděl ještě:

- metadata podmínek (*terms and conditions*): popisují podmínky využití objektu, právní omezení, omezení dané typem uživatele; dnes bychom tento typ řadili do administrativních metadat práv;

- administrativní metadata: vztahují se ke správě objektu na serveru nebo v repozitáři, např. datum poslední změny, datum vzniku aj.; dnes bychom tato metadata řadili také do metadat práv;
- metadata hodnocení obsahu (*content ratings*): popis vlastností objektu v konkrétním schématu pro hodnocení, které by přidělovala nějaká autorita, např. hodnocení vhodnosti obsahu pro děti; dnes bychom řadili do administrativních nebo popisných metadat;
- metadata o původu (*provenance metadata*): definují zdroj dokumentu a popisují jej včetně popisu aktivit a úprav na objektu; dnes se řadí do tzv. administrativních metadat nebo ochranných/technických;
- metadata vztahů a odkazů (*linkage or relationship data*): vyjadřují vztahy mezi různými objekty, např. více článků a číslem výtisku, které je obsahuje, mezi předlohou a jejím překladem apod.; dnes bychom považovali za strukturální nebo popisná metadata;
- strukturální metadata: definuje logické části složených objektů a přístup k nim; dnes se také nazývají strukturální metadata.

Opakujícími se skupinami v typologiích jednotlivých autorů jsou metadata popisná, administrativní, strukturální, technická, ochranná a metadata práv. Nejasnosti jsou kolem vztahu administrativních, ochranných, technických metadat a metadat práv. Tyto skupiny jsou někdy uváděny zvlášť, někdy je uváděno, že ochranná metadata a metadata práv jsou součástí metadat administrativních [NISO, 2004, s. 1].
V současné době je nejvíce přijímané rozdělení metadat na:

- popisná
- administrativní
    - metadata práv
    - ochranná metadata
    - technická metadata
- strukturální

Administrativní metadata jsou zastřešující skupinou pro metadata práv, ochranná a technická metadata, která se mohou různě prolínat. Viz také [CAPLAN, 2003, s. 3].

4.3.1 Popisná metadata

Popisná (někdy také deskriptivní) metadata popisují z bibliografického pohledu intelektuální obsah dokumentů, v našem případě jednoduchých nebo komplexních digitálních objektů. Jde tedy o typ údajů o autorovi, názvu, o fyzické předloze digitálního dokumentu, údaje o vydavateli apod. Jsou určena převážně pro uživatele a administrátory k vyhledávání, identifikaci dokumentů z jejich strany a výběr dokumentů. Kromě toho slouží také ke správě dat, vedle ostatních typů metadat. Z pohledu OAIS se jedná o *Popisnou informaci* (*Description Information*). Popisná metadata lze použít také k akvizici, sdílené katalogizaci, popisu vztahů popisované a jiné entity (např. u digitalizace, různá vydání stejného díla apod.). Pro popisná metadata, stejně jako původně pro knihy, platí tři známá pravidla Charlese A. Cuttera, která stanovil pro vytváření knihovního katalogu [CUTTER, 1889, s. 8]. Katalog má:

- umožnit člověku nalézt knihu pokud zná jejího autora, název nebo téma;
- ukázat jaké dokumenty má knihovna od konkrétního autora, ke konkrétnímu tématu a žánru;

- a pomoci s výběrem knihy podle vydání nebo tématu.

Pokud za pojem kniha, dosadíme digitální objekt, tak popis výborně funguje i pro digitální knihovnu, která může tyto body plnit právě s pomocí popisných metadat.

V posledních letech lze pozorovat stále podrobnější popisné záznamy, které již nepopisují pouze horní úroveň intelektuální entity jako je např. titul monografie, ale také jednotlivé její logické části (kapitoly, články) nebo fyzické části (stránky). Naopak menší pozornost je věnována věcnému popisu, snad díky přesvědčení, že tento typ popisu nahradí indexace a plnotextové vyhledávání. K tomuto se osvědčuje např. OCR v podobě ALTO XML, které je schopné text popsat až na úroveň znaků a výsledky zobrazit přímo v textu (resp. obrazu naskenované stránky, článku apod.). Pro digitální dokumenty se nejčastěji používají schémata Dublin Core, MODS, MARCXML, VRA, EAD aj. Metadata potřebná pro nalezení a následné zpřístupnění digitálního objektu nebo dokumentu lze do jisté míry považovat také za metadata ochranná [DAPPERT, 2009, snímek 23].

4.3.2 Administrativní metadata

Administrativní metadata slouží ke správě digitálních objektů a jsou určena administrátorům i automatizovaným systémům. Obsahují mj. technická metadata, metadata práv a ochranná metadata, která jsou většinou považována za samostatné podčásti administrativních metadat. Hranice mezi administrativními a popisnými metadaty není často zcela jasná. Mnohá popisná schémata obsahují několik elementů, které lze považovat za administrativní metadata (např. identifikátory, údaje o zpřístupnění, copyrightu aj.). Popisná metadata jsou většinou dostupná uživateli, zatímco administrativní spíše administrátorovi systému. Mohou obsahovat následující typ údajů: informace o akvizici, případně vstupu do archivního repozitáře; informace o právech pojících se k digitálním objektům; údaje o lokaci; technické vlastnosti objektu; informace o vhodném HW a SW prostředí pro zobrazení digitálního objektu aj. Synonymem pro administrativní metadata je standard PREMIS.

Z pohledu OAIS jsou metadata technická, administrativní a ochranná tzv. *Informací o reprezentaci* (*Representation Information*), která dává archivovanému objektu smysl a říká např., v jakém HW a SW lze objekt zobrazit, popisuje jeho historii, technické vlastnosti, dokumentuje autenticitu. V případě, že tyto informace chybí nebo nejsou dostatečné, je velmi těžké digitální objekt v repozitáři najít, spravovat, zobrazit nebo mu v budoucnu porozumět. Popisná a strukturální metadata jsou z pohledu OAIS *Popisnou informací* (*Description Information*). Je ovšem nutné říci, že toto rozdělení není pevné a některé části jednotlivých typů metadat jsou jak *Informací o reprezentaci*, tak *Informací popisnou*. Např. ochranná metadata popisují historii vzniku a životní cyklus digitálních dokumentů (*Description Information*), stejně jako údaje o formátu a HW a SW pro zobrazení objektů (*Representation Information*).

4.3.2.1 Technická metadata

Technická metadata popisují technické vlastnosti digitálních objektů a bitstreamu (méně již celých dokumentů). Cílem je uchovat údaje o technických vlastnostech digitálních objektů tak, aby s nimi mohly pracovat různé automatizované služby, aby bylo možné dle nich vyhledávat v archivním systému, filtrovat množiny konkrétních objektů s určitými vlastnostmi a provádět s nimi opatření dlouhodobé ochrany. Pokud chceme provádět např. migraci nebo jinou aktivitu týkající se

konkrétní verze obrazového formátu TIFF, musíme být schopni v repozitáři jednoduchým dotazem zjistit, jaké objekty jsou v tomto formátu uloženy. Technická metadata jsou typem metadat, která vznikají ve valné většině kompletně automaticky, pomocí různých metadatových extraktorů – viz kapitola 3.2.5.1. Některé typy údajů je možné aplikovat na všechny digitální objekty (kontrolní součet, údaje o datovém formátu), další pouze pro konkrétní typy (audio, video apod.). Technická metadata mohou mj. obsahovat následující typ údajů: informace o HW a SW, pomocí kterých byl digitální objekt vytvořen; informace o digitalizaci (formát, komprese, rozměry apod.); bezpečnostní údaje, jako jsou např. hesla, klíče; velikost souboru; kontrolní součet aj.
Z nejvíce používaných schémat v prostředí paměťových institucí jmenujme MIX pro obrazová data, audioMD pro zvukové soubory, videoMD[114] pro video soubory, případně hlavičky TEI nebo textMD pro textové digitální objekty[115]. Určitou množinu technických metadat obsahuje také PREMIS, část Object, který lze použít obecně na všechny typy digitálních objektů.

4.3.2.2 Ochranná metadata

Procesy logické dlouhodobé ochrany digitálních dat a *digital curation* jsou založeny na informacích, které doprovázejí data určená k ochraně. U digitálních objektů je menší šance rozpoznat a pochopit existující problémy. Nelze to většinou pouhým pohledem na digitální objekt, jako u fyzických knih. Je to ale možné pomocí zprostředkovaných informací připojených k digitálnímu objektu. Takovou informací jsou tzv. ochranná metadata. Procesům ochrany je potřeba poskytnout údaje popisující digitální objekt; údaje technické povahy k využití objektu a údaje popisující vše, co se během existence digitálního objektu s ním stalo. Ochranná metadata jsou dynamická, procházejí změnami neustále během životního cyklu objektu. Mění se s využíváním, ochranou a jakoukoliv jinou manipulací s popisovaným digitálním objektem. Tj. metadatový záznam, např. ve standardu PREMIS, je v repozitáři (LTP systému) neustále doplňován.
Obecně zajišťují ochranná metadata dva cíle. Prvním je poskytnutí potřebných údajů pro správce dat (repozitáře), který tak má možnost provést odpovídající opatření k logické ochraně digitálních objektů. Druhým cílem je zajistit, že obsah archivovaných objektů bude možné zobrazit a interpretovat v budoucnosti, a to i přes technologické změny.

Definicí ochranných metadat je několik. Nejpoužívanější je definice z *Datového slovníku PREMIS* [PREMIS EDITORIAL COMMITTEE, 2011, s. 3], který ochranná metadata definuje jako: „*informace, které repozitář používá k podpoře procesů dlouhodobé ochrany*“. Tato definice je velmi obecná a je tedy dále rozvedena. Skupina odborníků podílejících se na vytváření specifikace PREMIS vidí ochranná metadata jako podporu funkcí repozitáře vedoucích k zajištění životnosti, zobrazitelnosti, srozumitelnosti, autenticity a identity v kontextu ochrany. Ochranná metadata tak zasahují i do jiných kategorií metadat (administrativní, technická a strukturální). Naopak typické pro ochranná metadata je vyjádření údajů dokumentujících původ, historii objektu, interní i externí vztahy digitálních objektů a také HW a SW prostředí nutné pro zobrazení objektů.

[114] VIDEOMD, AUDIOMD i textMD schémata vznikla v Kongresové knihovně, která je dále udržuje, viz http://www.loc.gov/standards/amdvmd/ a http://www.loc.gov/standards/textMD/.
[115] TextMD schéma poskytuje např. elementy pro uložení a zápis údajů o koncích řádek, znakových sadách, pořadí bytů a také o technickém prostředí, ve kterém byl textový dokument vytvořen, nebo jej lze vytisknout/zobrazit. Zaznamenat lze také pořadí stránek, což se potom překrývá s údaji ve strukturálních metadatech v případě použití kontejneru METS.

Výstižnější se tedy zdá definice NISO z roku 2004, která popisuje ochranná metadata jako: *„formu administrativních metadat, která se zabývá údaji o provenienci zdroje a jeho správou v archiv.*" [NISO, 2004, s. 16] Když pracovní skupina OCLC/RLG publikovala v roce 2001 srovnání několika tehdejších specifikací ochranných metadat (CEDARS, NEDLIB a Národní knihovny Austrálie), shledala, že uvedené specifikace ochranných metadat mají společné cíle. Jedním z nich bylo: *„podporovat správu archivních objektů poskytováním relevantních informací správcům repozitář tak, aby mohli dělat rozhodnutí zajišťující přístup k obsahu dokumentů i navzdory měnícím se technologiím.*" [OCLC/RLG WORKING GROUP ON PRESERVATION METADATA, 2002, s. 18] Dalším společným bodem bylo, že specifikace se cíleně vztahovaly na digitální objekty obecně, bez závislosti na konkrétním formátu objektů.[116] Posledním průsečíkem v přístupu byla skutečnost, že ochranná metadata mají být použitelná pro jakákoliv opatření dlouhodobé ochrany.

Princip ochranných metadat vychází z OAIS (viz např. specifikace australské národní knihovny nebo výstupy skupiny OCLC/RLG z roku 2002, části standardu PREMIS, specifikace CEDARS a NEDLIB). Bylo totiž nutné a žádoucí vytvořit takovou high-level specifikaci ochranných metadat, která by byla obecně použitelná pro jakýkoliv typ dat, pro jakýkoliv typ ochranných procesů a na jakémkoliv archivačním systému. Díky OAIS referenčnímu rámci nebylo nutné vymýšlet další obecný koncept, OAIS bylo využito jako startovní bod, který byl v jednotlivých svých částech rozpracován do podoby konkrétních specifikací ochranných metadat. V mnoha specifikacích se tak informační model OAIS přímo odráží. Informace potřebná pro dlouhodobé uchování by měla být obsažena ve dvou Informačních objektech OAIS informačního modelu:

- *Informace o obsahu* (*Content Information*) – sestává většinou z informací o technické podstatě digitálního objektu; dává systému informace o možnostech zobrazení konkrétních formátů v konkrétních prostředích SW a HW; obsahuje údaje o velikosti souboru aj.; s migracemi se tato informace logicky musí měnit;
- *Popisná informace pro ochranu* (*Preservation Description Information*) – obsahuje další informace potřebné pro logickou dlouhodobou ochranu, správu a využití digitálního objektu, tj. identifikátory, bibliografické údaje, údaje o autorských právech, vlastníkovi, původu, historii změn a vztazích s ostatními objekty.

Právě tyto dvě části informačního modelu OAIS jsou rozpracovány ve většině specifikací ochranných metadat. Specifikace CEDARS používá i stejnou terminologii a názorně zobrazuje i jednotlivé podčásti *Informace o obsahu* a *Popisné informace pro ochranu*, tedy *context*, *provenance*, *reference* a *fixity information* – viz [OCLC/RLG WORKING GROUP ON PRESERVATION METADATA, 2001, s. 41-42]. Z tohoto konceptu vychází dnes nejrozšířenější schéma ochranných metadat PREMIS.

Je důležité si uvědomit, že bez metadat dokumentujících procesy vzniku objektu (např. úpravy obrazu u digitalizace; stažení/získání u digital-born dokumentů; údaje o vstupu do repozitáře), bez technických metadat (údaje o formátu), je obtížné sledovat stav digitálních objektů v repozitáři. Nelze udělat analýzu vedoucí k výběru digitálních objektů a jejich následném použití pro ochranné procesy (migrace např.). Ochranná metadata jsou díky sledování a zaznamenávání událostí spojených s objektem primárním zdrojem pro jeho zajištění autenticity a poskytují objektu

[116] Např. Australská národní knihovna ve své specifikaci ovšem vedle obecných elementů má elementy, které se použijí pro konkrétní typy digitálních objektů (např. audio, video apod.).

dokumentaci o něm samotném, jako součást AIP balíčku. Mohou obsahovat mj. následující údaje [NATIONAL LIBRARY OF AUSTRALIA, 2003, s. 94]:

- dokumentace o aktivitách provedených na ochranu digitálního objektu (migrace, obnovy);
- dokumentace o jakýchkoliv dalších aktivitách a změnách na digitálním objektu, které se vyskytly během uložení nebo vzniku digitálního objektu;
- informace o technickém prostředí potřebném pro zobrazení digitálních objektů (může být klíčové pro pozdější emulaci SW a HW prostředí);
- informace o technických vlastnostech objektu (identifikátor formátu z registru formátů);
- údaje o integritě, identitě a autenticitě objektů;
- údaje o právech duševního vlastnictví, omezeních použití;
- o kontextu digitálního objektu;
- aj.

Ochranná metadata nelze vytvářet ručně, je potřeba mít přizpůsobené systémy, které tato metadata vytvářejí a plní. To znamená, že např. systém workflow digitalizace by měl tato metadata generovat, podobně jako následně archivní repozitář, tj. LTP systém. Tento proces generování, ukládání metadat, správy událostí a změn je velmi náročný. Je tedy logické, že by se měl aplikovat opravdu pouze na dokumenty, které byly vybrány pro dlouhodobou ochranu na základě určitých kritérií. To se týká převážně tzv. archivních kopií a netýká se jakýchkoliv derivátů (náhledy, uživatelské kopie apod.).

4.3.2.3 Metadata práv

Metadata práv obsahují informace, které se týkají duševního vlastnictví spojeného s digitálním objektem nebo k SW/HW nutném pro jeho zobrazení (licence). Do metadat práv patří také údaje o povolení nebo omezení přístupů (administrativní práva) ke konkrétním digitálním objektům. Práva přístupů je nutno rozlišit pro digitální objekty určené pro běžného uživatele a pro objekty určené k dlouhodobé ochraně. Uživatelům jsou nabízeny uživatelské kopie, které se v LTP systémech většinou neukládají, nejsou určeny na dlouhodobou ochranu. Naopak je jasné, že budou v relativně krátké době nahrazeny jiným formátem. U uživatelských kopií jsou často práva přístupů a práva autorská nastavována v aplikaci zpřístupnění dle určitých pravidel (např. přístup z určité IP adresy, v závislosti na roce vydání dokumentu apod.). Výhodou tohoto řešení je, že v případě změny (např. autorský zákon) není nutno měnit všechna metadata digitálních objektů, ale pouze v aplikaci nastavené pravidlo. Tento přístup je plánován i v projektu NDK v aplikaci Kramerius4.

Druhým aspektem jsou metadata práv pro archivní digitální objekty. Může existovat potřeba zaznamenat údaje o duševním vlastnictví nebo omezení přístupů i pro ně. Pak se nejčastěji používá PREMIS část Rights a schéma METSRights[117], určené pro vložení do kontejneru METS. METSRights umožňuje vyjádřit převážně duševní práva (licence, vlastník i jeho kontaktní údaje, kontext apod.). PREMIS Rights dokáže vyjádřit podobné údaje, je ale flexibilnější. Dovoluje vložit své vlastní schéma s údaji o právech. Dalším způsobem řešení je, že jsou vytvořeny záznamy s metadaty práv, které mají v systému konkrétní identifikátor a jsou přes něj spojeny

[117] http://cosimo.stanford.edu/sdr/metsrights.xsd

s relevantními objekty. Pokud se něco změní, není nutno měnit metadata všech objektů, ale pouze záznamu s metadaty práv.

4.3.1 Strukturální metadata

Strukturální metadata popisují strukturu a vztahy mezi digitálními objekty tvořícími komplexní digitální objekt, nebo mezi různými digitálními objekty. Slouží jak k zobrazení uživatelské verze dokumentu, tak pro jeho správu.
Dnes je v digitalizaci periodik a monografií běžné skenování na úroveň stránky, to znamená, že např. z čísla periodika, které má 10 stran, vznikne 10 samostatných digitálních obrazů. Tyto obrazy nejsou přístupné jako smysluplný celek, pokud nejsou doprovázeny strukturálními metadaty, které lze využít v aplikacích zpřístupnění. Struktura může být vyjádřena dvojího druhu – logická a fyzická. Fyzická struktura váže jednotlivé stránky do „fyzického" celku, tj. říká, že těchto 10 stránek (digitálních objektů) tvoří dohromady jedno číslo periodika. Pokud má každá stránka více reprezentací (archivní kopii v JPEG 2000, uživatelskou kopii v JPG a OCR soubor v TXT), pak je to také vyjádřeno ve fyzické mapě. Druhým typem strukturálních metadat je tzv. logická struktura, která vyjadřuje logickou stavbu dokumentu, např. čísla periodika, které se skládá z článků jdoucích v nějakém sledu za sebou. U knihy se většinou vyjadřuje logická struktura na úroveň kapitol, příloh. Ve většině případů jsou strukturální metadata uvedena v obou typech, např. logická a fyzická strukturální mapa v METS záznamu.
Strukturální metadata jsou velmi důležitá pro zpřístupnění dokumentu jako celku, pro navigaci v něm a také pro dlouhodobé uložení, pokud archivní balíček AIP je uložen jako logický balíček.
V současnosti se v prostředí knihoven používá nejčastěji schéma METS a jeho část <structMap>. Existovalo také schéma EBIND, jako předchůdce METS. Vzniklo v roce 1996 na Univerzitě v Berkeley jako pokus o standardizaci strukturálních metadat vycházející z TEI DTD [CAPLAN, 2003, s. 160].

4.4 Vytváření metadat

Dlouhou dobu bylo vytváření metadat v rukou jedné komunity. Tou byli knihovníci a částečně také pracovníci ostatních paměťových institucí. Ti vytvářeli katalogizační záznamy pro fyzické dokumenty a sbírky, později s nástupem elektronického publikování i pro elektronické dokumenty. Nástup Internetu znamenal, že metadata začali vytvářet i neodborníci bez vzdělání katalogizátora, např. nakladatelé, autoři, samotní uživatelé a v prostředí digitalizace běžní pracovníci.
Vytváření metadat závisí na typu dokumentů, okamžiku kdy metadata vznikají a na typu workflow, ve kterém vznikají. Může probíhat automaticky nebo manuálně při vytváření digitálních objektů nebo v pozdějších částech jejich životního cyklu. Většinou ale jde o kombinaci obou přístupů.
Návrh metadat pro digitalizaci, který je součástí této práce, počítá s kombinací. Maximální množství metadat bude vznikat automatizovanou cestou, pomocí nástrojů na extrakci metadat. Další část se bude přebírat z katalogizačních záznamů. Analytický popis ovšem bude nutno dělat manuálně (popisy článků aj.).

Automatická tvorba metadat se týká v oblasti digitálních repozitářů převážně technických metadat, které jsou extrahovány metadatovými extraktory. Ty jsou schopné dodat v rámci tzv.

obohacení metadat v LTP systémech veškeré technické údaje o jakémkoliv digitálním objektu. Nástroje jako JHOVE1/ JHOVE2, NZME mají různé moduly pro různé typy digitálních objektů (TIFF, JPG, ARC aj.) – více viz kapitola 3.2.5). Automatická tvorba metadat je ideálem, kterému se snaží přiblížit veškeré projekty digitalizace nebo zpracování digitálních objektů. Automatizace je přesnější než manuální vytváření, rychlejší a tedy i levnější.

Automatickou cestou lze do jisté míry vytvářet i popisná metadata. Buď pomocí extrakce metadat ze souborů, které určitou sadu metadat již obsahují (MS WORD, PDF), nebo pomocí SW na rozpoznávání znaků, který je schopen dle specifického nastavení a algoritmů rozeznávat na naskenované stránce konkrétní oblasti (nadpis, místo kde je uveden autor apod.). Segmentace stránky na oblasti lze aplikovat nejlépe u dokumentů, které mají stejnou nebo podobnou strukturu (vědecké články), hůře se s ním pracuje v případě běžných periodik nebo monografií. I tak výsledky potřebují manuální opravu, pokud se nechceme smířit s určitou mírou chybovosti.[118]

Proběhlo několik pokusů, alespoň v teoretické rovině, vytvořit nástroj na automatické vytváření popisných metadat. Lze jmenovat nástroj MARS (*Medical Article Record System*) z Národní lékařské knihovny USA, který dokázal vyčítat na základě rozpoznávání bloků popisné údaje o článcích. Pokusů založených na segmentaci stránky bylo více – viz článek *Genre Classification in Automated Ingest and Appraisal Metadata* [KIM a ROSS, 2006].

Jednou z automatizovaných cest vzniku metadat jsou i konverze z jednoho schématu do jiného, většinou novějšího, nebo toho, který je používán v instituci, kam chceme poslat naše data spolu s metadaty. Na základě mapovacích tabulek lze vytvořit aplikaci, která dle převodních šablon bere hodnoty jednotlivých elementů původních záznamů a plní je do nového schématu. Dalším způsobem automatického vytváření metadat je tzv. harvestování (sklízení). Sklízeny jsou dostupné externí zdroje, které už potřebná metadata obsahují (katalogy, databáze apod.).

Manuální tvorba metadat využívá různých nástrojů, uzpůsobených většinou do šablon, které odpovídají používanému metadatovému schématu. Do těchto šablon pracovník dle určitých pravidel vyplňuje hodnoty elementů. SW na manuální tvorbu metadat se snaží pomoci a některé elementy plní automaticky, případně nabízí povolené hodnoty, provádí kontroly výsledku, zda jsou vyplněna povinná pole apod. Výsledek je nejčastěji XML záznam vygenerovaný aplikací tak, aby obsahoval vyplněné hodnoty a odpovídal používanému schématu metadat. Výhodou manuální tvorby je určitá kvalita za předpokladu, že pracovníci vědí, co dělají a mají potřebné vzdělání, dodržují dané standardy, pravidla a návody. Důležité je používání kontrolovaných slovníků, tezaurů a autorit. Celkově používání kontrolovaných termínů přispívá k interoperabilitě záznamů. Na druhou stranu je popsaný způsob manuální tvorby pomalý, časově a tedy i finančně náročný.

Další možností manuální tvorby metadat je zapojení samotných uživatelů, kteří v moderních aplikacích zpřístupnění mají možnost jednotlivé dokumenty komentovat, tagovat a také opravovat OCR záznamy (např. systém TROVE v Národní knihovně Austrálie[119]). Takovému využívání uživatelů k popisu nebo dalším úpravám se říká *crowd-sourcing* a je stále populárnější. Jedním ze

[118] Nizozemská národní knihovna digitalizuje od roku 2010 masovým způsobem periodika. Používá SW aplikaci na workflow digitalizace DocWorks od firmy CCS, které je schopno rozlišovat nadpisy, reklamy apod. I tak ovšem výsledky opravuje několik desítek lidí, aby bylo dosaženo stoprocentní přesnosti a správnosti určení nadpisů, reklam aj.

[119] Ukázka úprav OCR http://trove.nla.gov.au/ndp/del/article/12426381/1521841?zoomLevel=3.

zástupců je např. Captcha[120], kdy uživatel musí opsat text, aby se dostal k dalšímu kroku konkrétního úkonu, např. stažení souboru v prostředí Internetu. Finská národní knihovna vytvořila v rámci projektu DIGITALKOOT[121] jednoduché flashové hry, ve kterých k dosažení cíle musí hráč přepisovat jednotlivé texty a tím tedy vytváří/upravuje OCR soubor k jednotlivým obrazům. Podobné crowdsourcingu je také vytváření tagů, klíčových slov a podobných popisů, což je dnes tak oblíbené nejen ve webových službách typu YouTube, Flickr apod., ale také v digitálních knihovnách, které tuto možnost začaly v posledních letech nabízet. Například Kongresová knihovna od roku 2008 zpřístupňuje část svých digitálních sbírek fotografií i plakátů prostřednictvím serveru Flickr.com[122], kde uživatelé mohou doplňovat tagy, upřesňovat popisy jednotlivých dokumentů a také to dělají. K této problematice více např. [ZARRO a ALLEN, 2010]. Z uvedeného vyplývá, že kvalita těchto metadat může velmi kolísat a tvůrci málokdy dodržují jakákoliv pravidla vytváření záznamů, což může omezovat reálné využití těchto metadat.

Metoda tvorby metadat kombinující oba přístupy, automatický i manuální, je nejčastěji využívaná v prostředí digitalizace. Používají se specializované SW, které zvládají a spravují celé workflow tvorby digitálních objektů (např. DocWorks od firmy CCS). Pokud je využíváno standardní schéma metadat, je výběr mezi takovými nástroji relativně velký, úměrný použití standardu v komunitě. Typický postup tvorby metadat, pro který je navržen i metadatový profil popsaný v kapitole 7 a příloze této práce, vypadá následovně: Popisná metadata vrchních úrovní dokumentů (titul) jsou primárně přes OAI-PMH nebo Z39.50 protokol[123] stahována z katalogu nebo jiného zdroje. Ve většině procesů digitalizace platí nepsané pravidlo, že do procesu digitalizace je zařazen pouze dokument, který již je bibliograficky zpracován a má bibliografický popis/záznam – srovnej např. [UHLÍŘ, 2006, s. 4]. Důvodem je to, že v procesu digitalizace by se záznam musel vytvářet, což by zdržovalo. Následně musí být stažené záznamy čištěny (např. od údajů o lokaci z jiných knihoven), případně doplněny. Stažená popisná metadata procházejí migrací, např. z MARC21 do MODS v rámci workflow digitalizace. Další popisná metadata vznikají víceméně manuálně. Jde o vnitřní části dokumentu, ke kterým metadata v externích zdrojích neexistují (kapitoly, logické části, opravy číslování, zadávání názvů logických částí aj.). Dále se doplňují, ideálně automatickou cestou, administrativní metadata mapující jednotlivé úkony a kroky. Po zpracování obrazu dochází k finalizaci metadat, kdy jsou doplněny logické i fyzické struktury dokumentu a technická metadata k obrazu. Výsledek, různá metadata v různých schématech, vzniklá manuálně, automaticky, mohou být zabalena do kontejnerového schématu METS, který umožňuje zachovat jednotlivá schémata v jejich původní podobě v rámci jednoho METS záznamu. Výsledný XML záznam je uložen spolu s digitálním objektem, případně jsou nejdůležitější hodnoty elementů uloženy také v databázi systému. Velmi často vznikají různé množiny metadat pro archivní kopii dat a pro uživatelskou kopii, někdy i v různých formátech.

[120] http://www.captcha.net/
[121] http://www.digitalkoot.fi/fi/splash
[122] http://www.flickr.com/photos/8623220@N02
[123] Protokol pro vyhledávání informací v rámci knihovních aplikací; určen speciálně pro podporu vyhledávání z distribuovaných serverů; umožňuje provádět z aplikace na jednom počítači dotazy do databáze na jiném počítači.

4.5 Uložení metadat

Metadata mohou být uložena třemi různými způsoby. Mnoha datovým formátům je vlastní ukládat informace o sobě, svém vzniku, autorovi apod. přímo v souboru samotném. To platí např. pro JPEG 2000, obrazy JPG a jejich EXIF informaci, soubory MP3 a WAV, které mají metadata v hlavičce. I dokumenty MS Word nebo PDF obsahují vložená metadata. Firma Adobe vyvinula své metadatové schéma XMP (*Extensible Metadata Platform*), které používá na uložení metadat napříč svými aplikacemi (Reader, InDesign aj.). Někdy jsou takto uložená metadata nazývána „vlastnostmi“, ovšem z podstaty věci, tj. podávání informací o obsahu a samotném objektu, jde o metadata. Výhodou je, že metadata nelze od digitálních objektů oddělit nebo je ztratit. Problémem je ale jejich hromadné využití, indexace a úpravy. V případě webových stránek jsou metadata mnohdy součástí HTML kódu, např. v podobě Dublin Core.

Druhou možností je, že metadata jsou v podobě souboru (nejčastěji XML nebo např. TXT) uložena vně popisovaného objektu. Digitální objekt tedy neobsahuje metadata, ale obsahuje na ně přímý odkaz přes nějaký identifikátor. Takováto metadata, která jsou často součástí archivního balíčku spolu s digitálními objekty, je obtížné od digitálních objektů oddělit, což je výhodné z hlediska dlouhodobé ochrany. Zároveň na soubory metadat lze použít stejné nebo podobné metody ochrany jako na vlastní data. Jednoduchá je i indexace metadatového záznamu.

Třetí možností, z hlediska používanosti v repozitářích i z hlediska správy digitálních objektů nejčastější, je uložení metadat v nezávisle spravované databázi. Takováto metadata jsou propojena s objektem pomocí identifikátorů. Výhodou je rychlý přístup k metadatům, jednoduché vyhledávání, reportování. Databáze je schopna ukládat relační modely komplexních objektů. V databázi nejsou metadata uložena v syntaxi XML, ale jako jednotlivá pole databáze. Pokud je potřeba záznam sdílet, poslat, je nutno aby databáze dokázala vytvořit např. XML záznam odpovídající jednomu nebo více různým schématům metadat.

V současných LTP systémech je nejčastější formou uložení metadat uložení spolu s datovým (digitálním) objektem, kdy archivní balíček je samopopisný, což je jeden z cílů dlouhodobé ochrany i OAIS referenčního rámce. Zároveň jsou ale metadata uložena v databázi, kde se buď všechny nebo vybrané elementy metadat dublují s obsahem metadat uložených spolu s archivním objektem. Při úpravě metadat v databázi se musí změnit i záznam v archivním balíčku a naopak.[124] Tímto způsobem pracují LTP systémy a odpovídá referenčnímu rámci OAIS.

4.6 Metadata a XML

V posledních čtyřiceti letech došlo k posunu od specificky kódovaných (zapsaných) elektronických dokumentů k obecně kódovaným. Specificky kódovaný dokument obsahuje kódy, které určují, jak má být dokument formátován. Tyto kódy byly většinou typické pro konkrétní SW nebo kombinaci SW a HW a takto zapsaný digitální dokument nefunguje v jiném SW. Naproti tomu, obecné kódování (např. pomocí SGML, TEI, XML aj.) se zaměřuje na popis struktury a sémantického obsahu dokumentu, pravidla formátování nechává na externí dokumenty a nástroje. Tím je umožněno sdílet tyto dokumenty mezi systémy a různými prostředími. V 1969 na základě konceptu obecného kódování vznikl v IBM jazyk GML (*Generalized Markup Language*). V roce 1986 byla jeho vylepšená verze přijata za normu ISO pod názvem SGML [YOTT, 2005, s. 214].

[124] Nejčastěji používanou syntaxí v knihovním prostředí je MARC.

V roce 1996 vzniká na základě SGML specifikace XML. Struktura každého XML záznamu je definována v definici dokumentu – DTD (dnes XSD) souboru. Jde o soubor, který určuje, jaké elementy může XML záznam odpovídající určitému schématu mít, jak je omezen jejich výskyt, jak se mohou nořit do ostatních elementů a mnohdy určuje i možnosti plnění hodnot elementů.
XML samotné je pro metadata dnešní doby určující. Paměťové instituce byly první, které XML ve velké míře adoptovaly a začaly pro metadata používat. Výhodou byla flexibilita, indexovatelnost, přenos metadat a skutečnost, že XML záznam není „plochý“, ale hierarchický. Dokáže vyjádřit strukturu pomocí vnořených elementů a hierarchické vztahy mezi jednotlivými elementy záznamu. Navíc XML záznam je digitální objekt sám o sobě, což je rozdíl např. oproti MARC záznamu, který je pouze kontejnerem pro přenos popisných údajů v binární podobě. Na základě XML začaly vznikat nástroje a další návazné technologie, jako např. XSLT (*eXtensible Stylesheet Language Transformations*), XLink, XQuery, a další. XPath je způsob odkazování na část/element XML dokumentu; v kombinaci s XSLT umožní XPath extrakci hodnot elementů a převod do jiné podoby, např. pro zobrazení nebo další využití. Pokud s metadaty chceme dále provádět úpravy nebo je jinak využívat, XPath a XQuery jsou dvě metody, jak s nimi pracovat. Většina transformací mezi metadatovými schématy je prováděna pomocí XSLT šablon. Základní funkcí XSLT šablon je převod XML např. do HTML nebo jiného XML formátu. XSLT šablony využívají XPath pro zpracování XML záznamů. Využíváno je např. pro migrační tabulky a šablony. XSLT je tedy metodou, jak zobrazit údaje obsažené v elementech XML záznamu. XLink je způsob linkování mezi různými XML záznamy/dokumenty. XQuery je pokusem postavit obecný dotazovací jazyk pro XML dokumenty. Odlišnost od XPath je v tom, že na rozdíl od XPath XQuery nejde po částech XML a elementech ve stromovité struktuře, ale přistupuje k obsahu pomocí dotazů strukturovaného jazyka, podobně jako SQL [REESE, 2008, s. 91].
To, co dělá XML velmi užitečným, i když jde vlastně o vylepšený textový dokument, je právě prostředí, dostupné nástroje a technologie, které s ním souvisejí. Výhodou je také jeho „čitelnost“ člověkem, což neplatí např. pro MARC, pokud konkrétní osoba není odborník na MARC. XML je otevřený formát, který zaručuje, že se instituce nestává závislou na jedné technologii.

4.7 Konverze metadat

Konverze metadat je proces, kdy metadata zapsaná v jednom schématu je potřeba převést (konvertovat) do schématu jiného, cílového. Jde o mapování nejen elementů, ale i jejich sémantiky a syntaxe celého schématu.
Důvodů pro převod metadatového záznamu může být více, obecně jde vždy o zajištění interoperability mezi různými schématy. Nejčastějším důvodem pro mapování je používání novějšího schématu a tedy nutnost převodu všech starých metadatových záznamů. To může být spojeno s pouhou změnou standardu, nebo se změnou SW pro archivaci/zpřístupnění, který může podporovat jiné, než je stávající schéma metadat. Konverze se také často provádí při přijímání metadat do archivu nebo digitální knihovny, kdy příchozí metadata jsou převáděna do vnitřního standardu systému. Konečně důvodem konverze je i sdílení metadatových záznamů v různých systémech a agregace metadat v jednom systému pro vyhledávání. Mapování a vlastní převod je často zabudovanou vlastností konkrétních systémů, ať již na uložení nebo zpřístupnění dat. Ten příchozí metadata ve výchozím schématu systém automaticky převede na schéma chtěné.
I v případě, kdy jsou metadatová schémata, mezi kterými chceme udělat konverzi interoperabilní, tj. mají podobnou strukturu, mají elementy s podobným obsahem, je nutné vytvořit tzv.

sémantické mapování. Napřed v podobě *mapovací tabulky* (*crosswalk table*) a poté technicky toto mapování vyjádřit, např. jako XSLT šablonu. Mapovací tabulka obsahuje jednotlivé elementy výchozího schématu a k nim náležející (sémanticky) elementy schématu cílového. XSLT šablona konkrétnímu SW říká, jaký element a jaký atribut z původního metadatového záznamu odpovídá jakému elementu a atributu nového schématu a zda a v jaké podobě má přenést hodnotu původního elementu do elementu nového.

Existuje několik problematických oblastí, na které je dobré se při mapování a následné konverzi zaměřit [REESE, 2008, s. 159]. První z nich je konzistence výchozích metadat. Ta musí být jednotná a nemělo by docházet k tomu, že pomocí jedné konverzní tabulky budeme převádět různé verze byť stejného schématu. Každá specifikace metadat se během doby upravuje, mění se také často pravidla plnění hodnot elementů, jako tomu bylo např. v NK ČR pro data z VISK7, kde schémata DTD jsou ve třetí verzi a pravidla popisu jsou na tom podobně. Pokud existuje nekonzistence, výsledek konverze bude velmi špatný. Je nutné udělat konverze pro jednotlivé verze, nebo ještě lépe změn specifikací se v budoucnu zcela vystříhat.
Druhou problematickou oblastí je granularita. Velmi zřídka se převádí metadata mezi schématy, která mají stejnou úroveň granularity. Často je tomu přesně naopak, např. MARC21 do Dublin Core apod. V takovém případě se hledají elementy, které mají své protějšky, další elementy se posuzují podle toho, zda je možné jejich informaci vyřadit (nebude ve výsledném záznamu), nebo případně kam hodnotu výchozích elementů namapovat. V případě, že konvertujeme do stručnějšího schématu, přicházejí ke slovu řešení jako opakování elementů, používání elementů k uložení hodnot výchozí specifikace, která do nich sémanticky ne zcela patří apod.[125]
S granularitou jsou spojeny také problémy s povinností výskytu určitých elementů v cílovém schématu (konkrétní element nelze vynechat) a otázka různých počtů výskytů sémanticky stejných elementů (element je ve výchozím schématu opakovatelný, v cílovém schématu nikoliv). Je potřeba promyslet, zda půjde o tzv. mapování „jedna k jedné“, nebo bude potřeba hodnoty elementů upravovat, slučovat nebo naopak rozdělovat[126].
Třetí oblastí je rozhodnutí o případných ztrátách. Při mapování dvou schémat se většinou ukáže, že na obou stranách existují elementy, které nemají protějšek v druhém formátu. Nejde jen o ztráty hodnot některých elementů, ale často může jít o rozhodnutí vedoucí ke ztrátě kontextu nebo jedné celé části původního záznamu. Např. v EAD, které často popisuje jak archivní jednotku, tak archivní sbírku, se lze rozhodnout, že se nebudou mapovat údaje o archivní sbírce.
Konverze budou zcela určitě přesnější, pokud předchází analýza směřující k zjištění toho, jak široký je sémantický průnik obsahů jednotlivých elementů. Další podmínkou úspěšnosti je samozřejmě nejen shoda struktur obou formátů, ale také určité společné metodické postupy užité při naplňování struktur daty, tj. stručně řečeno společná metodika popisu dokumentů [KAŠPAROVÁ a PSOHLAVEC, 2008, s. 23]. Toto je ulehčeno tím, že v komplikovaných metadatových schématech je kladen důraz na kontrolovatelné slovníky a obecně na sémantiku hodnot elementů. Ta potom umožní mapování i v případech, kde sémantika dvou schémat je odlišná. Jde jen o to, najít vhodný protějšek pro konkrétní element.

[125] Příkladem může být umístění autora textu, autora ilustrace a autora předmluvy z MARC21 do elementu Creator v Dublin Core, který se bude opakovat bez bližšího určení role autora, která v MARC21 je zcela jasná.
[126] Možné varianty jsou *one-to-one*, *one-to-many*, *many-to-one*, *one-to-none*.

Mapování je vždy individuální a odvíjí se od lokální praxe. Vyžaduje spoustu rozmýšlení a rozhodování o tom, jak obě schémata spolu souvisejí a jakou mají vzájemnou podobnost, ať sémantickou nebo syntaktickou. Existují samozřejmě volně dostupné šablony pro konverze nejužívanějších schémat (např. z MARC21 do MODS, z MODS do Dublin Core a mnoho jiných), z nichž některé jsou i oficiální, tj. vytvořené např. autoritou, která standardy spravuje (Kongresová knihovna). Tyto volně dostupné konverzní tabulky a XSLT šablony jsou dobrým východiskem, ale je nutné je upravit na způsob použití formátů v konkrétní instituci. Samostatným problémem je vytváření konverzních tabulek pro proprietární standardy metadat, které jsou vytvořené a využívané pouze v naší instituci. V tomto případě žádné volně dostupné pomůcky neexistují a vše si musí instituce udělat sama. Tak tomu bylo i při přípravách konverzí DTD periodika a monografie do MARCXML (2007/2008) a později do MODS pro novou verzi aplikace Kramerius4 (2009-2010).

Za největší a nejviditelnější konverzi metadat v prostředí NK ČR můžeme považovat přechod ze standardu UNIMARC na standard MARC21. Přechodu předcházelo mnoho měsíců příprav a velmi pomohlo i to, že tyto standardy jsou si podobné a mají stejné zaměření. Podobná byla konverze záznamů DOBM SGML do DTD periodika a monografie a také do standardu MASTER+ v roce 2003 – více viz kapitola 5.2.

4.8 Vývoj metadat ve světě

Pokud přijmeme to, že za metadata lze považovat jakékoliv strukturované údaje vytvářené s cílem popsat konkrétní intelektuální entitu, pak lze říci, že metadata jsou s lidstvem od té doby, kdy lidé začali organizovat informace. V oblasti knihoven můžeme za metadata považovat papírové katalogizační záznamy, případně knih konkrétní sbírky vznikající už od starověku. Moderní katalogizační teorie a výstupy se objevily před sto padesáti lety. Důvodem bylo opět organizování informací pro vyhledávání v knihovnách [CHAPMAN, DAY a HIOM, 1998].
I přestože jsou metadata a katalogizační záznam založené na odlišném konceptu, hlavní cíl, popsat dokument, je stejný. Katalogizační záznam většinou obsahuje pouze popisné údaje o díle včetně určitých údajů o fyzické podstatě popisovaného dokumentu (vyjádření díla). Cílem katalogizačního záznamu je popis předlohy a umožnění vyhledávání uživateli. Záznam nemá žádný hlubší vztah k popisovanému dokumentu, není jeho součástí. Metadata naopak mohou být součástí digitálního objektu, který popisují (např. v hlavičce obrazového formátu), nebo jej alespoň doprovázejí, jsou s ním uložena. Metadata na rozdíl od katalogizačního záznamu podávají veškeré možné informace o digitálních objektech, včetně popisných. Jsou vytvářena tak, aby jim SW aplikace rozuměly a dokázaly s nimi dále pracovat. Jsou často vytvářena automaticky; v procesu digitalizace nebo samotnými autory/vydavateli, tedy pracovníky, kteří nejsou katalogizátoři a nemají ani ponětí např. o AACR2.
Z tohoto pohledu je katalogizační záznam podskupinou metadat a lze jej považovat za metadatový záznam. *„Ačkoliv se např. záznam v Dublin Core může zdát jako zjednodušený katalogizační záznam, je nutné si uvědomit, že kontext tvorby a využívání metadat je podstatně odlišný a vedený snahou překonat paradigma tradiční katalogizace. Považovat proces tvorby metadat za zjednodušenou katalogizaci by bylo nepochopení situace*.“ [GRADMANN, 1998] Podobně mluví i další autoři, metadata nejsou novým názvem pro katalogizaci nebo katalogizační záznam. Tradiční bibliografický popis je pouze malou částí velké množiny metadat pro podporu vyhledávání a správy. Katalogizační standardy běžné v knihovnách považuje za metadata také organizace NISO

[NISO, 2004, s. 1]: *„V knihovním prostředí, [termín] metadata je běžně používán pro jakékoliv formální schéma popisu zdroje, které lze použít na jakýkoliv typ objektu, digitálního nebo nedigitálního.*“ I přes toto konstatování se lze někdy setkat s názorem, že rozdíl mezi metadaty a katalogizačním záznamem spočívá v jejich formě. V elektronické podobě jde o metadata, pokud popis není v digitální formě, jde o katalogizační nebo jiný záznam. Někomu se právě toto zdá vhodné hledisko odlišení metadat od katalogizačního záznamu. Většina odborníků ale tvrdí, stejně jako my výše, že jde v obou případech o metadata [CAPLAN, 2003, s. 2]. Trefně to popsaly v článku *Metadata: Cataloging by Any Other Name* Jessica Milstead a Susan Feldman: *„Jako člověk, který roky psal poezii, aniž by o tom věděl, knihovníci a indexátoři produkovali a standardizovali po staletí metadata …*“ [MILSTEAD a FELDMAN, 1999]

Katalogizace až do 60. let 20. století probíhala pouze papírovou cestou. Záznamy se staly elektronickými až s příchodem standardu MARC (*MAchine Readable Cataloguing*). Záznamy v MARC formátech můžeme označit za metadata, popisují konkrétní intelektuální entitu a mají strukturovaný zápis.[127] Struktura je dána specifikací standardu MARC, obsah standardem pro plnění (např. AACR2) nebo kontrolovanými slovníky (např. LCSH aj.).[128] Za strukturovaný zápis bychom mohli prohlásit i záznamy podle ISBD (*International Standard Bibliographic Description*), které bylo publikováno ve své první verzi v roce 1974. Jeho přínosem byla předepsaná struktura záznamu. Ze záznamu je tak jasné, co která jeho část popisuje, aniž by uživatel musel například ovládat jazyk záznamu. ISBD lze považovat za předchůdce značkovacího jazyka [SMIRAGLIA, 2005, s. 5]. ISBD a AACR s rozvojem využívání digitálních dokumentů prošly změnami a revizemi. Vznikl *International Standard Bibliographic Description for Computer Files* (ISBD (CF)), později přejmenován na *International Standard Bibliographic Description for Electronic Resources Computer Files* (ISBD(ER)). Popis proměn samotného MARC standardu směrem k popisu elektronických publikací viz [CAPLAN, 2003, s. 62-64]. Pro knihovny je zapojení nových procesů a

[127] Počátky MARC standardu spadají do roku 1966, kdy Kongresová knihovna prováděla automatizaci svého katalogu a v souvislosti s tím se snažila upravit i proces katalogizace. Cílem bylo vytvořit standard pro elektronický záznam, který by reprezentoval bibliografické informace a dal se lehce využít na vyměňování těchto informací. MARC používá pole pevné i proměnné délky a je postaven na anglo-amerických katalogizačních pravidlech (AACR). MARC standardy jsou obecně velmi strukturované a sémanticky bohaté. Všechny současné standardy MARC odpovídají normám *Information Interchange Format* (ANSI Z39.2) a *Format for Information Exchange* (ISO 2709). Rodina tzv. MARC standardů se rozrůstala, např. UKMARC-Velká Británie, CANMARC - Kanada, USMARC - USA, IBERMARC – Španělsko, PICAMARC – Holandsko aj. Zvláštní postavení mají UNIMARC a MARC21. UNIMARC vznikl jako program organizace IFLA s cílem vytvořit standard schopný zajistit výměnu záznamů mezi všemi vyjmenovanými variantami. Byl poprvé publikován v roce 1977 pod názvem UNIMARC – Univerzální MARC formát. UNIMARC se používal i v ČR a v mnoha zemích celého světa. Na konci 90. let se všechny MARC mutace musely potýkat s integrací digitálních dokumentů do procesů knihoven a jejich katalogů. Tyto změny byly tak velké a náročné, že na ně rozvoj standardu UNIMARC nedokázal adekvátně reagovat. To a přidružené finanční problémy s podporou UNIMARCu zapůsobily u zemí váhajících s přechodem na tento formát a formát UNIMARC začal s novým tisíciletím ztrácet podporu na úkor nového formátu MARC21. Jeho výhodou oproti UNIMARCU bylo např. to, že byl vytvářen od počátku jako interní formát, na rozdíl od „výměnného“ UNIMARCu. MARC21 byl specifikován výlučně na plnění podle pravidel AACR2R a je tedy přesnější. Důvody proč se MARC21 stal de-facto mezinárodním standardem: *„jsou hlubší a jsou převážně velmi pragmatické: rostoucí kooperace a globalizace přinesly nutnost integrace a využití maximálního množství dostupných zdrojů, z nichž drtivá většina je ve formátu MARC 21*“ [STOKLASOVÁ, 2001]. Podrobnější popis rozdílů a dění okolo standardů na mezinárodní scéně viz [STOKLASOVÁ, 2004] a [STOKLASOVÁ, 2001].

[128] Standard MARC jako specifikace polí, která se plní, spolu s pravidly plnění (AACR2) lze dohromady považovat za specifikaci metadat, podobně jako např. Dublin Core, i když nejde o klasické metadatové schéma.

konceptů, které si vyžadují elektronické dokumenty a tvorba metadat, často bolestivé a obtížné. Je nutné si zvyknout, že doba jednoho univerzálního popisu pro všechny typy dat je minulostí. Klasická katalogizace měla dva hlavní cíle: 1) poskytnout bohatý bibliografický záznam včetně popsání vztahů mezi dokumenty; 2) umožnit sdílení záznamů a ušetřit tak práci katalogizátorům. AACR a MARC standardy těmto cílům vyhovovaly a vlastně je umožnily splnit. Bohužel AACR a MARC zklamaly na poli zpracování digitálních dokumentů v prostředí Internetu (údaje o právech, ochraně digitálních objektů, autenticitě apod.) [ZENG a QIN, 2008, s. 6].

Většího rozvoje a atraktivity pro širší technickou i knihovnickou obec se metadatům dostalo s rozvojem Internetu a možnostmi, které poskytoval při budování digitálních knihoven, sdílení digitálních dokumentů a vyhledávání v nich. Již v roce 1999 napsala Eva Bratková, že *„metadata hluboce souvisejí se vznikem a rozvojem sítě Internet a jejích služeb, především pak WWW. Zdá se, že právě v síťovém prostředí metadata nabírají zcela nové rozměry a význam …“* [BRATKOVÁ, 1999, s. 178] Podobně například také Zdeněk Uhlíř: *„Uvažování o metadatech se vynořilo především v souvislosti s rozšířením sítě sítí, s Internetem …“* [UHLÍŘ, 2002b, s. 84] Podle Evy Bratkové lze tedy rozlišit metadata vzniklá pro prostředí Internetu a metadata ostatní pro popis objektů v digitálních knihovnách před i po vzniku Internetu [BRATKOVÁ, 1999, s. 180]. Termín *metadata* se začal používat pro popisnou informaci, která doprovázela digitální objekty na Internetu. Bez metadat by digitální objekty nebyly vyhledatelné a mnohdy by bez dalšího vysvětlení bylo těžké pochopit jejich smysl a kontext (autor, datum vzniku apod.). Metadata pro dokumenty na Internetu také často měla za cíl umožnit vlastní dokument prohlížet, resp. pohybovat se v něm. Termín metadata se nadobro propracoval do knihovnické komunity po roce 1995, kdy vznikla specifikace Dublin Core. Dnes již můžeme doplnit, že rozvoj metadat souvisel také s rozvojem digitalizace jako takové i s rozvojem digital-born dokumentů, které stále ve větší míře od poloviny 90. let vznikaly a bylo je potřeba popisovat, stejně jako dokumenty digitalizované. Nejvíce schémat vzniklo v druhé půli 90. let 20. století. Schémata byla založena na SGML nebo později na XML a stále se jednalo převážně o schémata popisných metadat (EAD, Dublin Core, TEI aj.).

4.8.1 Popisná metadata v knihovnách (Dublin Core, MARCXML, MODS a TEI)

První reakcí na potřebu používat popisná metadata v prostředí digitálních knihoven, na Internetu a v různých systémech, bylo schéma Dublin Core. Schéma vzniklo v roce 1995 jako výstup snahy knihovníků, odborníků na Internet a SW vývojářů vytvořit sadu základních elementů tak, aby ji mohly sdílet všechny tři jmenované komunity a to pro popis jakéhokoliv objektu (digitálního, digitalizovaného a také fyzického). V době vzniku Dublin Core existovaly pouze dva druhy popisných metadat pro elektronické online zdroje. Prvním byly údaje, které shromažďovaly internetové vyhledávače (indexace), druhým byly klasické katalogizační záznamy, pokud vznikaly. Oba typy byly pro potřeby internetové komunity nevhodné a bylo nemožné s nimi popsat tak různorodý materiál, jaký se začal objevovat na webu. Metadata Dublin Core mohou být v hlavičkách HTML dokumentů[129] (jako HTML tag), nebo jako samostatný záznam doprovázející popisovaný objekt (v syntaxi XML nebo RDF).

[129] Ty se tak stávají samopopisnými.

Velká výhoda schématu, jeho jednoduchost, kterou představovalo 15 (původně 13) základních prvků popisu, se ukázala později v určitých případech býti i jeho slabinou. Ne vždy bylo možné pomocí základních prvků popsat všechny vlastnosti dokumentů, které bylo třeba. Dublin Core byl proto v roce 2000 obohacen o tzv. kvalifikátory, které umožňují upřesnit jednotlivé prvky popisu a tak je od sebe odlišit. Kvalifikátory, později označované jako „zpřesnění" (*refinements*), jsou dvojího typu: zpřesnění elementů (*element refinements*) a tzv. schémata zápisu (*encoding schemes*). Základních 15 prvků se od doby uvedení kvalifikátorů označuje jako „jednoduchý" Dublin Core, při použití kvalifikátorů jako „kvalifikovaný" Dublin Core. Výhodou může být i skutečnost, že Dublin Core nemá vlastní pravidla pro zápis hodnot, spoléhá se vždy na jiná podobná pravidla (kontrolované slovníky jako např. LCSH, MESH, MDT aj.), v elementech kde to může být potřebné. V roce 2003 se Dublin Core stal mezinárodní normou ISO 15836. Dublin Core není závislé na konkrétní syntaxi (např. na XML/SGML jako je TEI Header). Lze vyjádřit např. v HTML, v RDF i v XML. Všechny elementy Dublin Core jsou opakovatelné a nepovinné.
Dublin Core je dnes považován za nejvyužívanější schéma pro popis digitálních dokumentů, právě pro jeho jednoduchost. Je stále velmi často používán pro menší projekty, a to ve své kvalifikované podobě. Výhodou je i jeho flexibilita. Pokud v rámci konkrétního repozitáře naspecifikujeme vlastní rozšíření Dublin Core, jiné systémy, které pracují s Dublin Core si v takto upraveném záznamu najdou oněch patnáct základních elementů a záznam mohou dál využívat. Rozšíření tak nezpůsobuje nekompatibilitu. V současné době je Dublin Core využíván pro svou jednoduchost na zasílání metadat pomocí OAI-PMH, nebo na jednoduché vyhledávání v systémech digitálních knihoven přes protokol SRU/SRW nebo Z39.50. Je to také jedna z mála specifikací, na které se díky její flexibilitě dokáží shodnout systémy repozitářů[130], většina z nich ji podporuje. I přesto lze v současné době pozorovat odklon od Dublin Core, velké systémy a digitální knihovny častěji pracují s propracovanějšími schématy, jako jsou např. MODS, MARCXML a jiné.

Vzhledem k přílišné jednoduchosti Dublin Core a množině již hotových záznamů v MARC21 standardu vznikla na přelomu tisíciletí XML verze standardu MARC21 pod názvem MARCXML (někdy pod názvem MARC21XML). MARCXML je vlastně v XML syntaxi zapsaný MARC21. XML záznam obsahuje kódy polí i podpolí přesně, jak je definuje MARC21, tj. zápis v XML je neztrátový. Jediná odlišnost od klasického MARC záznamu je v podpoře kódování, které je pro MARCXML omezeno výhradně na UTF-8. MARCXML udržuje Kongresová knihovna a zajišťuje jeho podporu[131]. MARCXML navazoval na pokusy Kongresové knihovny vytvořit zápis MARC21 v SGML podobě. Koncem 90. let 20. století vzniklo DTD pro SGML a poté DTD pro XML aby obě DTD byla opuštěna ve prospěch nové specifikace MARCXML v roce 2002. MARC21 byl a stále je velmi zavedený standard pro popis dokumentů a jeho převod do XML byl snahou zaručit jeho využívání i v novém prostředí digitálních knihoven na Internetu. Při vzniku MARCXML bylo dbáno na to, aby schéma bylo jednoduché, flexibilní, rozšiřitelné a aby zaručovalo neztrátový převod z/do MARC21 záznamu. Výhodou byla možnost zobrazení pomocí XML šablon a možnost validace. I tak je možno napsat, že užívání MARCXML pro popis není rozšířené. Nejčastěji se používá jako prostředek převodu MARC21 do jiných schémat.
Problémem se ukázalo být to, že standard původně určený pro popis fyzických dokumentů ne vždy dokáže vyjádřit skutečnosti týkající se dokumentů digitálních, jejich vlastností a vztahů.

[130] DSpace má podporu pro kvalifikovaný Dublin Core přímo zabudovánu.
[131] http://www.loc.gov/standards/marcxml/

MARCMXL byl jednoduše řečeno příliš spjat s MARC21, ten se totiž z pohledu digitálních knihoven a digitálních dokumentů mnohým jeví jako archaický a nevyhovující. MARCXML pomáhá pouze zlepšit zacházení systémů s obsahem metadatových záznamů v MARC21 a vyjádření názvů elementů v číselných kódech, použití podpolí a indikátorů je zcela odlišné od ostatních konceptů běžně používaných metadatových schémat zapisovaných v XML.

I to byl jeden z důvodů, proč v roce 2002 vzniklo a bylo pro komentáře uvolněno schéma MODS (*Metadata Object Description Schema*[132]), jako jednodušší ale plně kompatibilní alternativa k MARCXML, určená a uzpůsobená navíc přímo k popisu digitálních dokumentů. Obě schémata koexistují dodnes s tím, že každé je zaměřeno jiným směrem. Cílem MODS bylo poskytnout alternativu k velmi jednoduchému Dublin Core, poskytnout možnost vyjádření metadat z existujícího záznamu MARC21 a také poskytnout schéma k tvorbě záznamů zcela nových. MODS není přesným vyjádřením MARC21 v jiné podobě, ale vychází ze stejné sémantiky. K zjemnění významů elementů používá běžné atributy, ne indikátory jako MARCXML. MODS zároveň nevyžaduje využití pravidel zápisu, jako jsou např. AACR2. Oproti MARCXML má dva hlavní rozdíly, prvním je slovní vyjádření elementů u MODS, namísto klasických kódů v MARCXML. Záznam samotný je tak lépe pochopitelný i člověku neznalému. Druhou odlišností je možnost libovolně přeskupovat elementy v rámci záznamu a tím vyjádřit strukturu popisovaného dokumentu. Lze tak popsat např. pouze číslo časopisu v jednom záznamu, ale také je možno vytvořit záznam časopisu, který obsahuje jednotlivá čísla, včetně popisu těchto čísel. Dublin Core, MARC21 a jeho XML podoba toto neumožňují. MODS je dnes používán i pro vytváření OAI-PMH profilů a stal se také podporovaným schématem popisných metadat ve spoustě systémů (např. Fedora, DSpace aj.). Je nutno poznamenat, že vývoj standardu MODS probíhal ve světle vývoje schématu METS, tedy tak, aby MODS bylo možno použít pro bibliografickou část METS záznamu [REESE, 2008, s. 132]. Takto je použit i v předkládaném návrhu metadat pro projekt NDK.
Výhodou MODS je nepochybně vazba na MARC21, který v knihovnách stále dominuje. Vazba přitom není tak silná jako v případě MARCXML. MODS tedy je kompatibilní se stávajícími bibliografickými záznamy, existují oficiální převodní tabulky. Proto je využíván často v digitalizaci, kdy jsou automaticky přebírány z katalogu záznamy v MARC21 a převáděny do MODS. Schéma je zároveň flexibilní, má dostatečnou granularitu elementů (19 hlavních elementů, spoustu atributů a dceřiných elementů). Pozdějším doplňkem k MODS je schéma MADS (*Metadata Authority Description Schema*). Jde také o XML schéma, které se používá k zaznamenání autoritních záznamů pro osoby, organizace, události a termíny. MODS není pouze pro zápis a uložení obsáhlých bibliografických metadatových záznamů, je vhodný také jako prostředník pro převody mezi MARC záznamy a záznamy, které nejsou v MARC formátu.

Na pomezí knihoven a humanitních, sociálních věd, lingvistiky a literatury stojí standard TEI (*Text Encoding Initiative*). TEI vzniklo v roce 1987 jako mezinárodní projekt na vývoj a následnou údržbu na HW a SW nezávislé metody zápisu digitálních humanitních a kulturních textů. Cílem bylo vytvořit návod na konzistentní zápis SGML digitálních textů a podnítit tak jejich využívání a výměnu odborníky v humanitních oborech [CAPLAN, 2003, s. 66]. Výstupem projektu bylo jedno z nejstarších metadatových schémat. Původně vycházelo ze SGML (TEI P3), nyní se již orientuje na

[132] http://www.loc.gov/standards/mods/

standard XML (od verze TEI P4). Nejdůležitějším výstupem TEI konsorcia[133] je sada návodů (*The Guidelines for Electronic Text Encoding and Interchange*)[134], které specifikují metody zápisu pro strojem čitelné texty [TEI CONSORTIUM, 2007].

Současná verze TEI P5 má 21 modulů, které lze libovolně využívat. *„Jedná se např. o modul „tei", který definuje třídy, makra a datové typy použité v ostatních modulech. Dále modul „core", který obsahuje deklarace elementů a atributů, které budou potřeba u popisu jakéhokoliv typu dokumentu a proto jsou doporučeny pro globální použití. Modul „header" poskytuje deklarace elementů metadat a atributů, ze kterých se skládá hlavička TEI, což je komponent nutný pro to, aby záznam odpovídal specifikaci TEI. Modul „textstructure" deklaruje základní strukturální prvky potřebné pro vyjádření struktury u většiny objektů, které mají strukturu podobnou knize. Tyto moduly tak většinou budou součástí většiny TEI schémat* [TEI CONSORTIUM, 2011b]. Dalšími jsou mj. moduly pro popis různých typů dokumentů, např.: „*verse*" pro popis veršovaných dokumentů; „*textcrit*" pro popis textové kritiky; „*namedates*" pro popis osob, jmen, míst a dat; „*figures*" pro popis tabulek, formulářů apod. Kompletní přehled viz [TEI CONSORTIUM, 2011b].

Velkou výhodou je, že schéma je možno použít i na popis různých dokumentů jako takových, nejen jejich plných textů. Umožňuje to tzv. hlavička TEI (*TEI Header*), ve které jsou uváděna bibliografická i jiná metadata popisující buď plný text, nebo digitální objekt. Pokud hlavička TEI chybí, neodpovídá záznam specifikaci TEI. TEI Header se skládá v klasickém TEI záznamu z elementů vrchní úrovně (částí) [TEI CONSORTIUM, 2011a]:

- <fileDesc> popis souboru – obsahuje plný bibliografický popis elektronického souboru, jde o nejpoužívanější část k popisu jakýchkoliv dokumentů;
- <encodingDesc> popis kódování – dokumentuje vztah mezi digitálním textem a zdrojem nebo zdroji, ze kterých vznikl (byl derivován);
- <profileDesc> popis textového profilu – poskytuje podrobný popis nebibliografických vlastností textu, hlavně použité jazyky, kontext a účastníky jeho vzniku;
- <revisionDesc> popis revizí – shrnuje historii změn souboru.

Z hlediska využití v NK ČR při popisu rukopisů a starých tisků je podstatná verze TEI P4, ze které vycházel standard MASTER, určený pro popis rukopisů a starých tisků. V současnosti využívaný standard metadat pro popis historických dokumentů vychází ze specifikace TEI P5 a byl vytvořen v evropském projektu ENRICH pod názvem ENRICH TEI P5 (enrich.xsd) – více viz kapitola 5.3.1. Standard ENRICH TEI P5 využívá vedle čtyř výše uvedených základních modulů navíc také moduly „*msdescription*", „*linking*", „*namesdates*", „*figures*" a „*transr*" [BURNARD, 2008].

4.8.2 Nové typy metadat (ochranná, technická, kontejnerová metadata)

Obrovský rozvoj digitálních repozitářů přinesl na konci tisíciletí zájem o nové typy metadatových schémat. Začala vznikat schémata určená pro technická, administrativní a jiná metadata, která byla nutná pro správu i dlouhodobou ochranu digitálních dat. Hybnou silou již nebyl bibliografický popis, ani Internet, ani sdílení digitálních objektů, ale zájem objekty dlouhodobě ochránit.

Ochranná metadata – více viz kapitola 4.3.2.2 – se začala objevovat již před rokem 2000 a to díky rozvoji povědomí a znalostí o logické dlouhodobé ochraně digitálních dat a také díky rozvoji systémů na uložení dat. Podklad pro vznik svébytné skupiny metadat poskytl referenční rámec

[133] Konsorcium TEI vzniklo v roce 2000 a nadále schéma udržuje a rozvíjí.
[134] http://www.tei-c.org/Guidelines/

OAIS, který byl publikován v roce 1999 ve své první verzi. Mezi prvními, kdo publikoval své představy o ochranných metadatech a jednotlivých elementech byla v roce 1998 americká pracovní skupina *RLG Working Group on Preservation Issues of Metadata***.** Cílem skupiny bylo vytvořit specifikaci metadat pro zdigitalizované dokumenty, která by napomohla jejich dlouhodobé ochraně. Vznikla sada šestnácti elementů [RESEARCH LIBRARIES GROUP, 1998], mezi nimiž je sice element <changeHistory> (historie změn) nebo <validation> (validace), ostatní elementy se ale z dnešního hlediska zdají býti spíše technickými metadaty (typ skeneru, údaje o barvě, kompresi, kontrolních tabulkách apod.). Podstatné je, že již tato první specifikace ilustruje rozdílné zaměření ochranných a popisných metadat. Neobsahuje totiž žádné elementy popisné, vhodné pro vyhledávání. Metadata jsou striktně určena pro správu digitálních objektů (údaje administrativní a strukturální). Problémem této první specifikace je přílišná orientace na konkrétní typ digitálních objektů, kterým jsou digitalizované dokumenty. Není obecná a nebylo možné ji použít na jiné typy dokumentů.

Tuto specifikaci vzali mj. do úvahy v Austrálii, kde v roce 1999 Colin Webb z australské národní knihovny vytvořil první vyspělou obecnou specifikaci ochranných metadat [WEBB, 1999]. Pomocí této specifikace bylo možno popisovat různé typy digitálních dokumentů (obrazy, audio, video, text, databáze i spustitelné soubory), jednotlivé objekty i jejich sbírky. Jednotlivé elementy bylo možné aplikovat na dílo jako virtuální entitu, na jeho manifestaci v podobě komplexního digitálního objektu a také na jednotlivé jeho části, digitální objekty (soubory).[135] Za klíčovou autor považoval schopnost udržovat v systému uložení, a tedy i v metadatech digitálního objektu, veškeré záznamy o změnách, které se s digitálním objektem děly během jeho životního cyklu.

Specifikace měla 25 elementů. Mezi 25 elementy jsou mimo jiné datum vzniku objektu, informace o SW a HW potřebném pro jeho zobrazení, údaje o uložení, validaci, vztazích k jiným objektům a také informace o rozhodnutí k archivaci a o procesech, které se s objektem děly. Tato specifikace byla mj. předlohou pro specifikaci novozélandskou.

RLG specifikaci si vzali jako vzor také v britském projektu CEDARS[136] (*CURL Exemplars in Digital ARchiveS*), v jehož rámci vznikla v roce 2000 další specifikace ochranných metadat *Metadata For Digital Preservation: The Cedars Project Outline Specification* [RUSSELL, et al., 2000]. Ta byla podstatně podrobnější než šestnáct elementů ze skupiny RLG. V rámci projektu byl vyvíjen systém na archivaci a ukázalo se, že je potřeba vytvořit specifikaci metadat, které jsou důležité pro logickou ochranu jakéhokoliv typu digitálního dokumentu bez závislosti na zvolené metodě dlouhodobé ochrany. Specifikace CEDARS vznikala v úzké vazbě na referenční rámec OAIS a stejně jako výstup pozdější pracovní skupiny OCLC/RLG (viz níže), se zabývá dvěma částmi *Informačního objektu*, jak jej definuje OAIS, a to *Informací o obsahu* (*Content Information*) a *Popisnou informací pro ochranu* (*Preservation Description Information*). K těmto entitám definuje jednotlivé relevantní elementy. Je popsán tedy jak vlastní obsah (struktura, způsob zobrazení, sémantická informace aj.), tak i doprovodná ochranná informace k němu (odkazy, kontext, původ objektu, důvod ochrany, původní technické prostředí potřebné pro zobrazení, údaje o provedených změnách, metadata o autorském právu, licencích, údaje o integritě aj.). Specifikace neřešila otázku granularity, tj. nebylo určeno, zda a jaké elementy se vztahují např. ke komplexnímu digitálnímu objektu jako celku nebo k digitálnímu objektu (souboru). Kompletní přehled elementů viz [RUSSELL, et al., 2000, s. 11-31].

[135] Odpovídá přístupu pozdějšího standardu PREMIS.

[136] http://web.archive.org/web/20030306105927/http://www.leeds.ac.uk/cedars/index.html

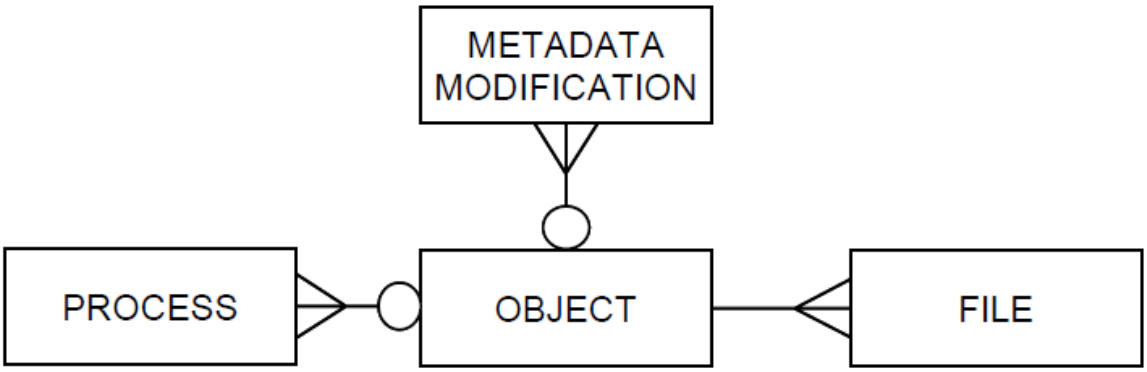
Podobně významným počinem v oblasti ochranných metadat byla specifikace vzniklá v rámci projektu NEDLIB (*Networked Europian Deposit Library*) [LUPOVICI a MASANÈS, 2000]. Stejně jako specifikace CEDARS vycházel otevřeně z informačního modelu OAIS, včetně terminologie a specifikace high-level typů informací, ze kterých se skládá archivní balíček AIP. Cílem projektu bylo navrhnout archivní systém DSEP (*Deposit System for Electronic Publications*; více viz [VAN DER WERF-DAVELAAR, 1999]) a bylo nutné pro něj specifikovat minimální množinu ochranných metadat. Vznikla tak tedy specifikace pro následující části OAIS: *Informace o reprezentaci* (*Representation Information*), která obsahovala např. údaje o HW a SW pro zpřístupnění, operačním systému, formátu digitálního objektu; a pro *Popisnou informaci pro ochranu* (*Preservation Description Information*), která obsahovala mj. identifikátory, kontrolní součet, údaje o změnách a další [LUPOVICI a MASANÈS, 2000, s. 18-25]. Návrh ochranných metadat NEDLIB obsahoval 8 elementů nejvyšší úrovně a 38 dceřiných elementů.
Všechny výše zmíněné specifikace měly podobný základ, inspirovaly se navzájem a vznikaly ve stejném období, po vzniku referenčního rámce OAIS, který se ukázal býti velmi vhodným výchozím konceptem i pro další rozvoj ochranných metadat.

Po roce 2000 aktivity ve vývoji dalších specifikací ochranných metadat pokračovaly v USA, na Novém Zélandu i jinde. Přelomem byl rok 2002, kdy byly publikovány specifikace pracovní skupiny OCLC/RLG, Národní knihovny Nového Zélandu a také specifikace ochranných metadat v rámci OCLC pod prostým názvem *Digital Archive Metadata Elements* [OCLC, 2002]. Specifikace byla součástí dokumentace k archivnímu systému, který OCLC vyvíjelo a nabízelo knihovnám. Specifikace obsahuje 34 elementů. S OCLC ochrannými metadaty je možné popsat mj. SW a operační systém nutný pro použití digitálního objektu, obsah digitálního objektu, jeho tvůrce, identifikátory, údaje o události a datum události spojené s životním cyklem digitálního objektu, popis změn funkcionality digitálního objektu, datum uložení do archivu, velikost digitálního objektu, datový formát[137] a jiné [OCLC, 2002]. Koncept byl nápadně podobný jako u pozdějšího standardu PREMIS (první verze v roce 2005), ovšem OCLC popis nebyl tak propracovaný. Jeho poslední revize proběhla v roce 2004.

Specifikace vznikající v Národní knihovně Nového Zélandu pod názvem *Preservation Metadata: Metadata Implementation Schema* [NATIONAL LIBRARY OF NEW ZEALAND, 2003], byla do té doby nejpropracovanějším návrhem ochranných metadat. Autoři se snažili najít rovnováhu mezi podrobnými nároky na metadata popsanými v referenčním rámci OAIS a mezi reálnou použitelností a možnostmi tvorby metadat. Specifikace vznikala s přihlédnutím k podobným aktivitám v OCLC/RLG v USA [OCLC/RLG WORKING GROUP ON PRESERVATION METADATA, 2002] i v Národní knihovně Austrálie [NATIONAL LIBRARY OF NEW ZEALAND, 2003, s. 4]. Byla určena pouze pro tzv. archivní kopie. Ostatní deriváty archivních kopií, jako např. uživatelské kopie a náhledy, nejsou považovány za objekty hodné logické dlouhodobé ochrany a není nutno je tedy popisovat ochrannými metadaty.[138]
Datový model novozélandské specifikace má čtyři entity: *Procesy*, *Objekty*, *Soubor* a *Změna metadat*. *Objekt* je definován jako archivní kopie (digitální objekt); *Proces* je definován jako jakákoliv aktivita provedená na archivním objektu; *Soubor* je entita obsahující technickou

[137] Pouze údaj typu „PDF“, v roce 2002 neexistovaly registry formátů, jako jsou např. UDFR a PRONOM dnes.
[138] Tento přístup se drží ve většině případů dodnes, uživatelské kopie, náhledy apod. nejsou předmětem logické dlouhodobé ochrany, pouze zálohování, tedy ochrany bitstreamu.

informaci o nejnižší úrovni digitálního objektu, tedy o souboru; *Změna metadat* je entita definovaná jako změna nebo úprava existujícího metadatového záznamu. Datový model znázorňuje Obrázek 7 níže.



Obrázek 7 – Datový model ochranných metadat [NATIONAL LIBRARY OF NEW ZEALAND, 2003, s. 6].

Z obrázku vyplývá, že *Objekt* může mít asociován jeden nebo více *Procesů*; jednu nebo více *Změn metadat* a také jeden nebo více *Souborů*. *Proces* musí vždy být spojen s jedním *Objektem*[139], podobně jako *Soubor* a *Změna metadat*. Elementy jednotlivých entit jsou podobné jako u specifikace OCLC, ovšem více podrobnější – viz přehled [NATIONAL LIBRARY OF NEW ZEALAND, 2003, s. 30].

- Entita *Objekt* mj. obsahuje datum vzniku, identifikátory, nutný SW a HW pro zobrazení objektu, strukturu (složený, jednoduchý objekt), logické členění, tvůrce metadatového popisu a datum jeho vzniku.
- Entita *Proces* popisuje mj. identifikátor relevantního objektu, důvod vyvolání procesu, kdo ho vyvolal, na základě jakých povolení, jaký byl výsledek procesu a jaké měl kroky, datum ukončení procesu.
- Entita *Soubor* má profily popisu podle typů dokumentů (audio, video, text, obraz, dataset, systémový soubor).[140] Každý z nich obsahuje základní technická metadata.
- Entita *Změna metadat* obsahuje elementy popisující identifikátor objektu, datum změny, informace o změněných datech a polích.

Dalším významným dokumentem publikovaným roku 2002 je zpráva zpracovaná ve spolupráci OCLC a RLG pod názvem *Preservation Metadata and the OAIS Information Model: a Metadata Framework to Support the Preservation of Digital Objects* [OCLC/RLG WORKING GROUP ON PRESERVATION METADATA, 2002]. Právě tato zpráva měla vliv na vývoj ochranných metadat v národních knihovnách Nového Zélandu, Austrálie, Německa i jinde. Zároveň dala základ standardu PREMIS, který později (po roce 2005) v oblasti ochranných metadat naprosto převážil.
Zpráva vznikla v rámci pracovní skupiny vytvořené již v roce 2000 v rámci organizací OCLC a RLG. Cílem bylo vytvořit obecnou specifikaci ochranných metadat aplikovatelnou na různé typy digitálních dokumentů a ochranných aktivit. Ihned na počátku práce této skupiny odborníků logicky vznikla *state-of-the-art* zpráva přinášející definici ochranných metadat, dále shrnující použití ochranných metadat ve světě a porovnávající dostupné existující specifikace [OCLC/RLG WORKING GROUP ON PRESERVATION METADATA, 2001]. Konkrétně šlo o specifikace z projektů

[139] Naproti tomu v PREMIS specifikaci lze proces/událost spojit s více Objekty.
[140] Na rozdíl od PREMIS, který je specifikací obecnou pro všechny typy digitálních objektů.

CEDARS a NEDLIB (*Networked Europian Deposit Library*) a také z australské národní knihovny. Všechny byly porovnávány mezi sebou (jednotlivé elementy) a bylo vytvořeno také mapování na relevantní části referenčního rámce OAIS.

Specifikace ochranných metadat v závěrečné zprávě vycházela z referenčního rámce OAIS a jejím účelem bylo vytvoření implementace informačního modelu ochranných metadat pro knihovnickou komunitu, která řeší otázky logické dlouhodobé ochrany digitálních dat. „*Informační model OAIS představuje high-level popis typů informací vytvářených a spravovaných komponentami archivního systému. Nezáleží na typu spravovaných digitálních objektů ani na technologii. Model OAIS poskytuje užitečný základ pro specifikaci ochranných metadat pro velmi široké použití.*“ [OCLC/RLG WORKING GROUP ON PRESERVATION METADATA, 2002, s. 9] Bylo ovšem nutno některé jeho části rozpracovat a specifikovat z pohledu reálného využití. V rámci pracovní skupiny proto vznikla rozpracovaná struktura OAIS a sada elementů ochranných metadat namapovaná na jednotlivé informační koncepty a specifikace OAIS modelu. AIP balíček v pojetí OAIS se skládá ze čtyř typů *Informačních objektů*, každý z *Informačních objektů* se skládá z vlastního obsahu (*Content Data Object*) a z *Informace o* (jeho) *reprezentaci* (*Representation Information*) – více viz kapitola 4.9.1. Pracovní skupina OCLC/RLG rozpracovala dva *Informační objekty* OAIS, které se přímo týkají ochranných metadat. Jsou to *Informace o obsahu* (*Content Information*) a *Popisná informace pro ochranu* (*Preservation Description Information*). U každého tohoto *Informačního objektu* byly navrženy elementy pro popis *Informace o reprezentaci* (*Representation Information*).

Návrh sady metadat staví na již existujících specifikacích OCLC, australské národní knihovny a projektu CEDARS (*CURL Exemplars in Digital ARchiveS*). Z každého ze jmenovaných projektů byly použity nějaké elementy, které byly doplněny elementy novými. Struktura celé specifikace a jednotlivé elementy již velmi předznamenávají návazný standard PREMIS (jeho část PREMIS Object a PREMIS Event). Oproti PREMISu ovšem zcela chybí zmínky o právech a o tzv. Agentech (osobách, SW nebo institucích), které stojí za nějakou událostí provedenou na archivovaném digitálním objektu. Kompletní seznam jednotlivých elementů a vysvětlující schémata jsou uvedeny ve zprávě [OCLC/RLG WORKING GROUP ON PRESERVATION METADATA, 2002, s. 48-50].

Aktivita pracovní skupiny OCLC/RLG vyústila ve vznik další skupiny odborníků, kteří v roce 2005 vydali první verzi standardu PREMIS – více viz kapitola 7.2.3. PREMIS měl od počátku ambice stát se jednotícím mezinárodním standardem v záplavě různých specifikací ochranných metadat. Druhým cílem bylo poskytnout tzv. jádro ochranných metadat, které by bylo možné použít na vytvoření ochranných metadat pro jakýkoliv digitální objekt, na kterém se plánují provádět nebo již provádějí *jakékoliv* metody dlouhodobé ochrany, a který je uložený v jakémkoliv typu repozitáře. Jinými slovy, co potřebuje jakýkoliv repozitář (jeho systém) vědět, aby byl schopen provádět aktivity dlouhodobé ochrany a zajistit zpřístupnění, pochopení apod. pro uživatele budoucnosti.

O standardu PREMIS dostala česká knihovní komunita informace již v roce 2005, díky příspěvku Martina Vojnara na konferenci *Archivy, knihovny a muzea v digitálním světě* [VOJNAR, 2006, s. 61].

Vývoj ochranných metadat se ale neodehrával pouze v anglo-americkém světě. Významným příspěvkem do celkového snažení je německá specifikace LMER (*Long-Term Preservation Metadata for Electronic Resources*). Standard vznikal v německé národní knihovně od roku 2003 a to ještě před publikováním standardu PREMIS. Důvodem bylo, že žádný finální vyzkoušený a

všeobecně přijímaný standard pro ochranná metadata neexistoval. Pracovníci německé národní knihovny si proto pomohli a vytvořili standard LMER, který používají dodnes. První verze LMERu (v. 1.0) vznikla v roce 2004 a byla pouze interní. LMER vznikal s vědomím, že probíhají práce na podobném standardu v USA v rámci pracovní skupiny OCLC/RLG *Preservation Metadata Framework*. Výsledek práce skupiny, PREMIS Data Dictionary, byl publikován počátkem roku 2005, stejně jako poslední aktualizace standardu LMER 1.2 z dubna 2005 [STEINKE, 2005, s. 4].

Německá národní knihovna používá LMER ve svém LTP systému DIAS, který byl mj. navržen tak, aby se standardem uměl pracovat. Záznam ochranných metadat vzniká při vytváření digitálních objektů a dále je doplňován v nástroji na Ingest (koLibRi[141]). Schéma se používá také v projektu národní sítě LuKII (*LOCKSS und KOPAL Infrastruktur und Interoperabilität*)[142] na ochranu digitálních dokumentů, postavené na aplikaci LOCKSS[143] (2010-2012). Nasazení standardu PREMIS se plánuje, až německá národní knihovna bude mít nový LTP systém (snad v roce 2013) [STEINKE, 2011].

LMER vychází ze specifikace ochranných metadat Nového Zélandu z roku 2002, záměrně je ovšem jednodušší [STEINKE, 2011]. LMER obsahuje jen základní elementy, stejně jako novozélandský standard, aby byl možný popis různých typů digitálních objektů. I to byl důvod, proč od počátku bylo možné k LMERu připojit jakákoliv jiná metadata ve formátu XML do elementu <xmlData> [STEINKE, 2005, s. 5]. LMER, stejně jako první verze PREMIS, se skládá z více XSD souborů (hlavní LMER[144] XML; LMER Object[145]; LMER Process[146]; LMER File[147] a LMER Modification[148]). Jednotlivé části obsahují:

- <lmerObject> – elementy popisující digitální objekt (počet souborů, kontrolní součet, název, datum vzniku apod.);
    - <lmerProcess> – elementy dokumentující všechny technické změny provedené na každém digitálním objektu; lmerProces může být vnořen do elementu <lmerObject> nebo <lmerFile>; údaje o změně obsahují např. datum změny, SW provádějící změnu, jméno osoby provádějící změnu aj.;
    - <lmerFile> – elementy popisující jednotlivé soubory digitálního dokumentu, např. velikost, název, kontrolní součet, MIME type, vazby na další soubory, aplikaci ve které soubor vzniknul; obsahuje element <xmlData>, do kterého je možno vložit jakýkoliv jiný XML stream; lmerFile je vnořen do části lmerObject;
- <lmerModification> – elementy popisující změny vlastního LMER záznamu, včetně data, aplikace, typu modifikace.

Obsah elementů je tedy podobný jako ve standardu PREMIS, spousta údajů ovšem oproti PREMISu chybí. Odlišná je ovšem struktura, kde např. v PREMISu je část PREMIS Agent a Event oddělena a lze na ni linkovat z PREMIS Object. Ve standardu LMER je opačný přístup, události jsou součástí popisů většiny částí, nejsou odděleny. LMER neobsahuje metadata práv. Podrobnější popis včetně jednotlivých XSD viz [STEINKE, 2009] nebo [STEINKE, 2005].

[141] http://kopal.langzeitarchivierung.de/index_koLibRI
[142] http://www.d-nb.de/wir/projekte/lukii.htm
[143] http://www.lockss.org/lockss/Home
[144] http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:1111-2005041114
[145] http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:1111-2005041208
[146] http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:1111-2005041212
[147] http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:1111-2005041220
[148] http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:1111-2005041239

Podobně nastal rozvoj technických metadat, která jsou pro obrazová data specifikována např. ve standardu MIX. Technická metadata pro jakýkoliv typ digitálního objektu mohou být mj. součástí standardu PREMIS Object a mohou ve velké většině vznikat automaticky pomocí nástrojů, jako jsou JHOVE, DROID, FITS aj. Technická metadata jako taková samozřejmě existovala již předtím, např. ve fotografii, kde při digitálním snímkování vznikají metadata a jsou součástí vlastního obrazového souboru (např. EXIF, XMP od firmy Adobe aj.). Více o standardu MIX v kapitole 7.2.4.

Po roce 2000 začaly vznikat také tzv. kontejnerové standardy, které se velmi rozšířily a jsou používány ve spoustě oborů, nejen v paměťových institucích. Dokáží pojmout (zabalit) různé typy metadat, např. administrativní, strukturální, popisná aj. Ve velké většině jsou doporučeny metadatové standardy, které je vhodné do těchto kontejnerových schémat plnit, je ovšem možné využít jakékoliv schéma i mimo doporučení. Kontejnerové standardy vyjadřují také strukturu popisovaného objektu/dokumentu včetně názvů a lokací souborů, které digitální objekt, nebo intelektuální entitu, tvoří. Vyjadřují tak vztahy mezi jednotlivými soubory jednoho komplexního digitálního dokumentu i jeho vztahy s jinými digitálními objekty. Kontejnerové standardy a záznamy v nich vzniklé jsou ideální na uložení a výměnu archivních balíčků, které odpovídají OAIS a jsou na to často používány. Je tak zcela jedno, jaký typ digitálního objektu je popisován, kontejnerový standard je obecný rámec, který umožní popis všech typů objektů.
Ze standardů, které mají silné zastoupení a využívanost v paměťových institucích lze jmenovat standardy METS, MPEG-21, v rámci vývoje SW pro digitální repozitář Fedora[149] vzniklý standard FOXML (*Fedora Object XML*), případně také standard pro balíčky vznikající archivací webu WARC (*Web ARChive*). Ten vznikl jako nástupce standardu ARC v rámci IIPC[150] (*International Internet Preservation Consortium*) v roce 2005 a od roku 2009 je mezinárodní ISO normou číslo 28500:2009[151]. Mezi další metadatové kontejnery patří např. OAI-ORE (*Open Archives Initiative Object Reuse and Exchange*) pro popis a výměnu agregovaných webových zdrojů, *XML Formatted Data Unit*[152] (XDFU) aj.
Nejrozšířenějším „kontejnerem" je standard METS – více viz také kapitola 7.2.1. Vznikl na popud americké *Digital Library Federation* (DLF[153]) a je udržováno a vyvíjeno mezinárodní radou (*METS editorial board*) pod dozorem Kongresové knihovny. METS je přímý nástupce projektu *Making of America II*. (MOA2), který měl za cíl vytvořit standard pro zápis popisných, administrativních a strukturálních metadat ke konkrétnímu digitálnímu dokumentu, který není knihou ani seriálem[154]. Celé úsilí bylo mířeno na digitální knihovny, zpřístupnění a výměnu takovýchto záznamů. V roce 2001 vzniklo MOA2 DTD [CANTARA, 2005, s. 238]. Bylo to právě v době, kdy byl publikován v první verzi referenční rámec OAIS. OAIS popisuje koncept a strukturu tří informačních balíčků pro přenos, archivaci a zpřístupnění. MOA2/METS lze využít pro všechny tyto balíčky – viz kapitola 3.3.2. Specifikace MOA2 integrovala popisná, administrativní a strukturální metadata do jednoho záznamu XML tak, aby bylo možno uložit a exportovat jeden záznam s různými metadaty uvnitř, namísto několika souborů XML, které popisují nebo se jinak pojí ke konkrétnímu digitálnímu dokumentu. Později byla specifikace přejmenována na METS. METS byl míněn jako výměnný

[149] http://fedora-commons.org/
[150] http://netpreserve.org/about/index.php
[151] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=44717
[152] http://sindbad.gsfc.nasa.gov/xfdu/
[153] http://www.diglib.org/
[154] Těm se věnovala první fáze projektu MOA1 od roku 1997.

formát, ovšem hledisko balícího mechanismu pro uložení dat a metadat později výrazně převýšilo jeho využití na výměnu dat. Podle výzkumu pracovní skupiny RLG a OCLC pro ochranná metadata z roku 2004, byl METS nejpoužívanějším standardem pro balení různých typů metadat a vyjádření ne-popisných metadat [OCLC/RLG PREMIS WORKING GROUP, 2004, s. 45].

4.8.1 Metadata v ostatních paměťových i jiných institucích

Vedle typických knihovnických formátů vznikaly i standardy metadat v oblastech mimo knihovny. Aktivní v oblasti metadat bylo a stále je archivnictví. V roce 1984 byl vytvořen standard MARC AMC (*Archival and Manuscripts Control*), který je dnes včleněn do MARC21. MARC AMC umožňoval popsat archivní sbírku pomocí MARC záznamu. Šlo tedy do jisté míry o popis tzv. archivní pomůcky. V roce 1994 byl publikován dodnes využívaný *General International Standard Archival Description* (ISAD (G)) vytvoření Mezinárodní archivní radou (ICA).
Archivnictví je úzce spjato s tzv. *recordkeeping* systémy (systémy pro správu záznamů). Metadata pro tyto systémy se vyvíjela dříve než standardy metadat pro archiválie, a to od 90. let 20. století. Později vznikly např. specifikace *Australian Recordkeeping Metadata Schema*[155] (RKMS); v Národním archivu Velké Británie doporučení pro použití metadat u ukládaných záznamů *Requirements for Electronic Records Management Systems*[156] apod. Na oblast se zaměřily také projekty, jako např. InterPARES[157] (*The International Research on Permanent Authentic Records in Electronic Systems*) v roce 1999[158].
V roce 1998 vznikl dodnes v archivnictví nejrozšířenější standard EAD[159] (*Encoded Archival Description*). V první verzi bylo EAD založeno na značkovacím jazyce SGML, šlo tedy o ze SGML odvozené DTD. V současnosti samozřejmě existuje verze XML s EAD XML DTD. Schéma je určeno k popisu a vytváření tzv. archivních pomůcek v elektronické podobě a také k popisu archiválií. Hlavním cílem vývoje bylo zpřístupnění archivních pomůcek (inventáře, soupisy apod.) v prostředí Internetu v elektronické podobě; schopnost schématu zachovat hierarchické vztahy mezi různými úrovněmi popisu; možnost pohybovat se po stromové struktuře informace v záznamu. Hlavní popisnou entitou je archivní fond, ne dokument jako je tomu např. v knihovnách. Práce na EAD, tehdy pod jiným názvem, odstartovaly v roce 1993 v Berkeley (Kalifornská univerzita) na základě požadavku Společnosti amerických archivářů. Již tehdy existovaly možnosti využít HTML, ale HTML jakožto SGML DTD neumožní vyjádřit rozdíly mezi hodnotami elementů, sémantiku. Vzniklo proto nové DTD ze SGML určené přímo pro zápis archivních pomůcek. Daniel Pitti a jeho tým specifikovali elementy tak, aby nové schéma dokázalo popsat většinu reálných papírových pomůcek, které pro tento účel nashromáždili. V roce 1995 tým z Berkeley publikoval schéma FINDAID DTD ke komentářům. Po změnách v roce 1998 byla publikována první verze EAD DTD, již pod tímto názvem. Tato verze byla postavena na XML [THURMAN, 2005, s. 187]. Standard je udržován Kongresovou knihovnou, což mu dává záruku dalšího rozvoje. EAD určuje sémantiku popisu archivní pomůcky, ale původně neposkytovalo žádný standard pro plnění hodnot

[155] http://www.infotech.monash.edu.au/research/groups/rcrg/projects/spirt/deliverables/austrkms-schemes.html
[156] http://www.nationalarchives.gov.uk/documents/metadatafinal.pdf
[157] http://www.interpares.org/
[158] InterPARES pokračuje dodnes a je ve své třetí fázi. Cílem první fáze bylo vytvoření teorie a metod k zajištění dlouhodobé ochrany autenticity záznamů vytvořených a udržovaných v databázích a systémech na správu dat (*document management systems*).
[159] http://www.loc.gov/ead/

elementů. Až později začalo vycházet v určitých bodech z TEI a ISAD(G). EAD DTD má tři sekce a tedy tři elementy nejvyšší úrovně: hlavičku <eadheader>, která obsahuje informace o záznamu samotném včetně základních informací o archivní pomůcce (např. ID); podrobnější popis archivní pomůcky včetně jejího úvodního textu <frontmatter>; popis obsahu a kontextu archivního fondu nebo sbírky <archdesc>. Hlavní popis je soustředěn na sbírku jako takovou, tj. její obsah, zaměření a administrativní informace. Druhá úroveň popisu se zaobírá logickými skupinami dokumentů, poslední úroveň pak individuálními dokumenty samotnými.

Vedle EAD vzniklo v roce 2001 také schéma EAC (*Encoded Archival Context*), která jde dál než EAD a dovoluje zapsat informace o autorech a kontextu vytváření archivního materiálu, který pomůcka popisuje. Údaje o autorovi je možné vyjádřit i v EAD, v elementu <bioghist>, ale ukázalo se jako užitečnější mít oddělený a formalizovaný záznam.[160]

V ČR se EAD schéma, na rozdíl od anglo-saských států v čele s USA, zatím masově neujalo. Např. v projektu *Ad Fontes*, který probíhá v Archivu hlavního města Prahy, se používá vlastní (proprietární) standard IDA AMP v.2.0[161], který obsahuje popis digitalizace i vlastní archiválie [HANOUSEK, 2010, s. 34]. V českém archivnictví se od roku 2006 používá tzv. *Standard pro ukládání a zasílání archivních pomůcek druhu inventář a dílčí inventář v digitální podobě*, který vytvořil Odbor archivní správy při Ministerstvu vnitra ČR. Pro matriky existuje obdoba – *Standard pro jednotnou evidenci matrik a výměnný formát pro ukládání a zasílání záznamu matrik v digitální podobě* [ČESKO. MINISTERSTVO VNITRA. ODBOR ARCHIVNÍ SPRÁVY, 2010]. V rámci projektu *Možnosti a formy zpřístupnění archivních fondů nebo jejich součástí veřejnosti v elektronické podobě*, který probíhal v letech 2007-2009 v Národním archivu ČR, vznikla mj. analýza shody českého standardu pro ukládání archivních pomůcek a EAD a také obsáhlé překlady relevantní k EAD, jako např. *The EAD Cookbook – 2002 Edition*[162]. *„Z analýzy vyplynulo, že český technický standard OAS MV je velmi kvalitní a vzhledem k tomu, že již při jeho vzniku se přihlíželo k standardu EAD, bude možné EAD implementovat.*“ [NÁRODNÍ ARCHIV, 2009, s. 2] Český standard bude ponechán a vedle něj paralelně vznikne standard odpovídající EAD. Snad i proto v zadávací dokumentaci projektu NDA (Národní digitální archiv), publikované na podzim 2011, jsou obsaženy požadavky na podporu schématu EAD, samozřejmě vedle jiných standardů.

V oblasti muzejnictví byla a je snaha po výměně popisných metadat a standardizaci menší než v knihovnách, ale i tak vznikl např. formát pro popis a výměnu záznamů muzejních objektů (*CIMI – Computer Interchange of Museum Information*; v současnosti je využíván minimálně).

Metadata se také ukázala potřebnými ve státní správě. Od roku 1994 byl vyvíjen americký standard GILS[163] (*Government Information Locator Service*) pro popis digitálních dokumentů. Zajímavostí je, že GILS je napsán jako aplikační profil protokolu Z39.50 pro vyhledávání v katalozích. Měl napomoci vyhledávání vládních dokumentů v různých systémech. Podobně ve Velké Británii vznikl např. e-GMS[164] (*e-Government Metadata Standard*).

[160] Důvody vzniku EAC a jeho cíle jsou podobné jako v prostředí knihovnických autorit, tj. mít jeden autoritní záznam, sdílet tyto záznamy tak, aby nebylo nutné je znovu a znovu vytvářet. Cílem bylo i sdílení s knihovnickými autoritami, kde archiváři mají více materiálů k upřesnění než samotní knihovníci.

[161] http://www.ahmp.cz/ida/v20/ida_amp_v20.xsd

[162] http://www.nacr.cz/Z-files/moznosti_08.pdf

[163] http://www.gils.net/

[164] http://interim.cabinetoffice.gov.uk/govtalk/schemasstandards/metadata.aspx

Pro oblast vzdělávání vznikla např. schémata GEM[165] (*Gateway to Educational Materials*) pro popis výukových dokumentů na webu a IEEE LOM[166] (*Learning Object Metadata*) pro popis objektů v digitálních výukových systémech. LOM má komplikovanější strukturu než většina běžných schémat a je schopno zaznamenat popisná, administrativní a technická metadata, včetně metadat práv – více viz např. [ZENG a QIN, 2008, s. 44-48].

Popis vizuálních objektů jako jsou obrazy, sochy apod. má silné zástupce ve standardech VRA Core[167] (*Visual Resoucrce Association*) a CDWA[168] (*Categories for the Descriptions of Works of Art*). Fotografie, obrazy, umělecké objekty byly zpočátku popisovány pomocí MARC standardu a pravidel AACR2, která se v kapitolách 8 a 10 těmto typům dokumentů věnují. Tento způsob popisu nebyl zcela odpovídající a nepokryl všechna specifika. Vznikla proto speciální metadatová schémata. Schéma CDWA vznikalo na počátku 90. let 20. století z popudu odborníků z oblasti muzeí, galerií a historiků umění. Cílem schématu je poskytnout rámec pro mapování popisných údajů ze stávajících systémů a pro vývoj systémů nových. Tedy specifikovat obecné popisné elementy pro popis uměleckých děl. CDWA zároveň obsahuje slovníky a postupy popisu, které mohou zajistit lepší kompatibilitu a přístupnost informací v různých informačních systémech věnovaných umění [ZENG a QIN, 2008, s. 35]. Jinými slovy CDWA je nástroj jak vyjádřit obsah databáze popisem informací o uměleckých dílech, které tyto databáze obsahují. Z tohoto důvodu je schéma velmi rozsáhlé, musí postihnout ideálně všechny možnosti popisu různých typů objektů, a obsahuje skoro 400 elementů. První verze CDWA vyšla v roce 1994. Schéma předepisuje sémantiku a pravidla popisu, nepředepisuje však syntax. Počítá se spíše s tím, že data jsou držena v databázi spíše než v záznamu. Později vznikla i verze *Lite* pro poskytování metadat přes protokol OAI-PMH.
VRA Core, nyní ve verzi 4.0, byl vyvinut jako odpověď na potřeby popisu sbírek s „vizuálními" zdroji v oblasti knihoven, především univerzitních. Ty bylo těžké popsat, protože obrazy, fotky, prezentace často nemají název, tvůrce a další údaje, používané v klasickém popisu v prostředí knihoven. Často naopak mají popis od původce, který se těmto knihovním přístupům vymyká [ZENG a QIN, 2008, s. 39]. VRA vychází ze standardů CDWA a Dublin Core. Oba rozšiřuje tak, že lze popsat umělecká díla, fotografie, budovy nebo např. sochy. Mimo to umožňuje také popsat objekty, které tato díla popisují nebo jinak znázorňují, jako je např. dokumentace ve formě mikrofilmů nebo digitálních obrazů. VRA Core vzniklo v roce 1997 a je od roku 2007 jedním z doporučených metadatových standardů pro plnění do kontejnerového standardu METS, části popisných metadat <dmdSec> [VISUAL RESOURCES ASSOCIATION, 2007]. Standard v nedávné době vzala pod svá křídla Kongresová knihovna, která nyní zajišťuje jeho podporu[169]. Schéma není tak rozsáhlé jako CDWA, má pouze 19 hlavních elementů, z nichž každý jen pár dceřiných elementů a atributů[170]. VRA je podobně ploché schéma jako např. Dublin Core.
Z komerční oblasti nakladatelů je nejznámější standard ONIX[171] (*ONline Information Exchange*), pro výměnu popisů publikací, které jsou nabízeny online mezi distributory, vydavateli, prodejci

[165] http://www.cen-ltso.net/main.aspx?put=215
[166] http://ltsc.ieee.org/wg12/index.html
[167] http://www.vraweb.org/projects/vracore4/
[168] http://www.getty.edu/research/publications/electronic_publications/cdwa/cdwalite.html
[169] http://www.loc.gov/standards/vracore/
[170] Ve verzi 3.0 byly tyto vnořené elementy vyjádřeny jako kvalifikátory, podobně jako v Dublin Core, jen v propracovanější podobě. Důvodem změny bylo mít VRA kompatibilní s XML.
[171] http://www.editeur.org/8/ONIX/

apod. Cílem bylo sjednotit způsob popisu těchto publikací na Internetu a zajistit tak jejich výměnu mezi systémy. Standard byl vytvořen samotnými vydavateli a obsahuje proto také údaje navádějící uživatele ke koupi dokumentu, včetně informací obchodního charakteru pro samotné vydavatele a distributory. ONIX lze namapovat na MARC21. Z pohledu knihovníků je ONIX schéma zvláštní v tom, že obsahuje podobné elementy jako tradiční katalogizační záznam, má ale rozdílná pravidla plnění a hodnoty těchto elementů mohou být pro knihovníky nezvyklé.

Pro oblast popisu audia a videa jsou nejznámějšími zástupci schémata MPEG-7 (*Multimedia Content Description Interface*) a MPEG-21 (*Multimedia Framework*). MPEG-7 definuje metadatové elementy, strukturu a vztahy, které používá pro popis audiovizuálních objektů včetně obrazových souborů, grafik, 3D modelů, hudby, videa a multimediálních sbírek. MPEG-21 byl vyvinut kvůli potřebě obecného rámce, který by zajistil interoperabilitu digitálních multimediálních objektů [NISO, 2004, s. 8]. Má 18 částí, z nichž se používají jen některé. Ty, které se používají, jsou postupně schvalovány jako mezinárodní normy[172]. Zvláště část 2[173] (*Digital Item Declaration*) je využívána a to i v projektech digitalizace. Nizozemská národní knihovna v Haagu ve svém projektu na digitalizaci 8 milionů stran novin, který začal v roce 2008, dodnes využívá MPEG21-DIDL (*Digital Item Declaration Language*) jako kontejner pro zabalení různých typů metadat a dat. Může takto postupovat díky tomu, že základní stavebním kamenem MPEG-21 standardu je digitální jednotka (*digital item*). *„Digitální jednotka je kombinací zdrojů (video a audio stopy, obrazy aj.), metadat (popisná, identifikátory aj.) a struktury (popis vztahů mezi zdroji). Druhá část MPEG-21 formátu specifikuje jednotné a flexibilní schéma pro interoperabilitu, které deklaruje strukturu a vytváření digitálních jednotek. Ty jsou deklarovány pomocí Digital Item Declaration Language (DIDL). Deklarace digitální jednotky spočívá ve specifikaci jejích zdrojů, metadat a vzájemných vztahů.“* [ISO/IEC 21000-2:2003, 2005]

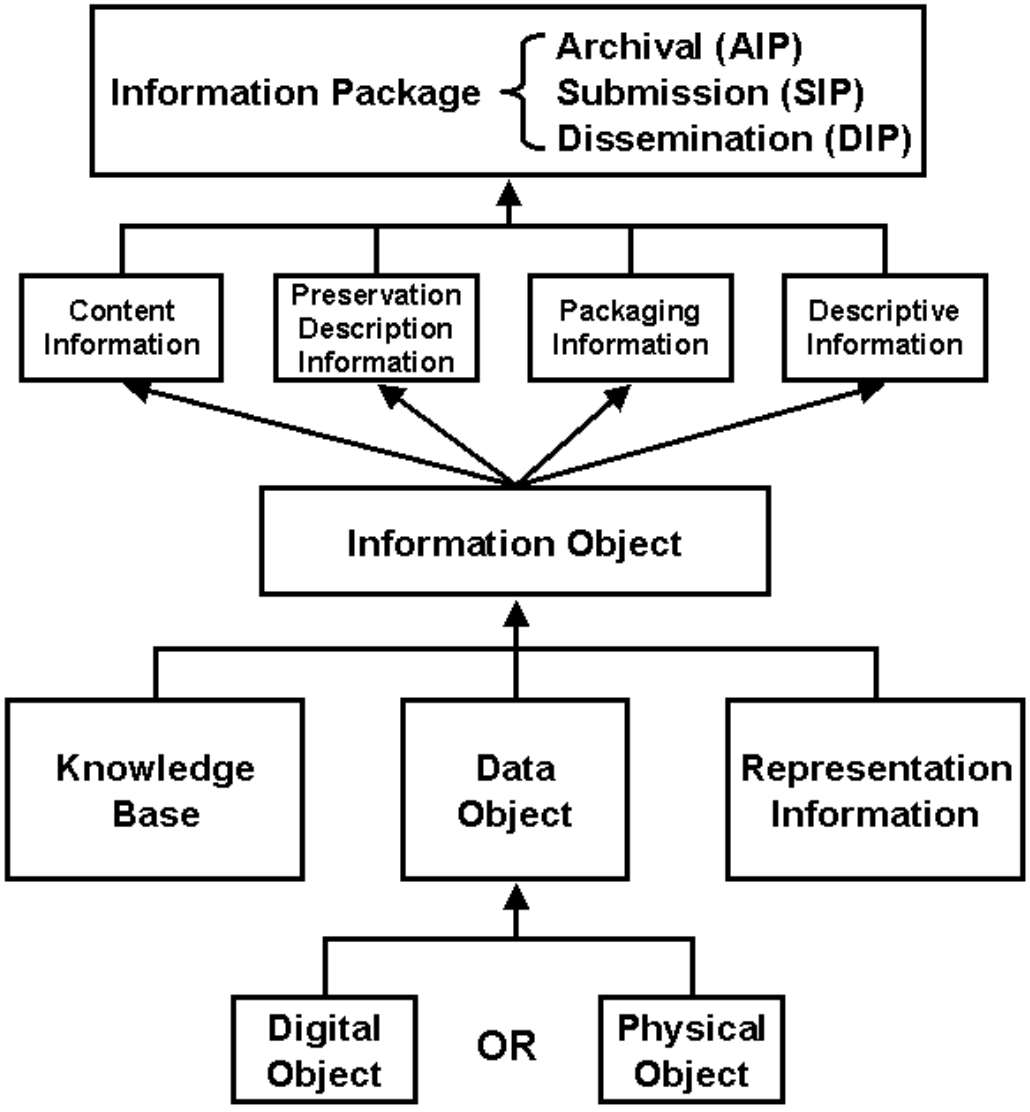
4.9 OAIS referenční rámec a metadata

Od svého vydání v roce 1999 se referenční rámec OAIS stal výchozím bodem pro každé seriózní budování digitálního repozitáře. Ukázala to již zpráva pracovní skupiny OCLC a RLG pro ochranná metadata, která byla vydána v roce 2004 a mj. konstatuje, že podle OAIS postupuje většina paměťových institucí při budování svého repozitáře a bere jej za výchozí bod i ve vztahu k metadatům a jejich specifikaci [OCLC/RLG PREMIS WORKING GROUP, 2004, s. 26-27]. Při vzniku OAIS bylo přihlíženo ke zprávě *Preserving Digital Information* [GARRETT a WATERS, 1996], která byla v mnoha ohledech přelomová. V této zprávě byly specifikovány vlastnosti ovlivňující schopnost zajistit dlouhodobou ochranu dokumentu a jeho informační integritu. Byly to obsah (*content*); celistvost (*fixity*); reference; provenience (*provenance*) a kontext. Tvůrci OAIS tyto vlastnosti zakomponovali beze změny do informačního modelu OAIS, jako čtyři části *Informace pro dlouhodobou ochranu*, která je vedle *Informace o obsahu* a *Informace o balení* součástí *Informačního balíčku* dle OAIS. V referenčním rámci OAIS se mluví o informacích, v reálném světě jsou ovšem tyto vyjádřeny metadaty. Podoba metadat tak vychází z OAIS konceptu tří *Informačních balíčků* (SIP, AIP, DIP – více viz kapitola 3.3.2). Většina organizací, které se otázkou dlouhodobého uchovávání digitálních dokumentů seriózně zabývají, tento model tří balíčků implementuje, podobně jako dostupné SW systémů pro správu repozitáře nebo LTP systémů.

[172] http://www.iso.org/iso/iso_catalogue/catalogue_ics/catalogue_detail_ics.htm?csnumber=30819
[173] http://www.iso.org/iso/iso_catalogue/catalogue_ics/catalogue_detail_ics.htm?csnumber=35366

4.9.1 Informační model OAIS

S jednotlivými digitálními objekty musí být uchován nejen informační obsah uchovávaných objektů, ale také další informace o původu a historii změn dokumentu, o jeho kontextu a zdrojích potřebných k porozumění. Informační model OAIS znázorňuje Obrázek 8 níže. V kontextu OAIS může informace existovat ve dvou formách, buď jako fyzický objekt (tj. např. papírový dokument, vzorek půdy), nebo jako digitální objekt (např. PDF soubor, TIFF soubor). Tyto dva druhy – fyzický a digitální – jsou obecně známy jako tzv. *Datové objekty*. Takový *Datový objekt* může nabývat různých forem, ať je to materiál, který je cílem ochrany, nebo metadata asociovaná s archivním digitálním objektem [OCLC/RLG WORKING GROUP ON PRESERVATION METADATA, 2002, s. 5]. Archivní objekt a jeho metadata jsou logicky dva různé datové objekty, i pokud jsou metadata součástí objektu (např. HTML dokument s metadaty).



Obrázek 8 – Informační model OAIS [OCLC/RLG WORKING GROUP ON PRESERVATION METADATA, 2002, s. 6].

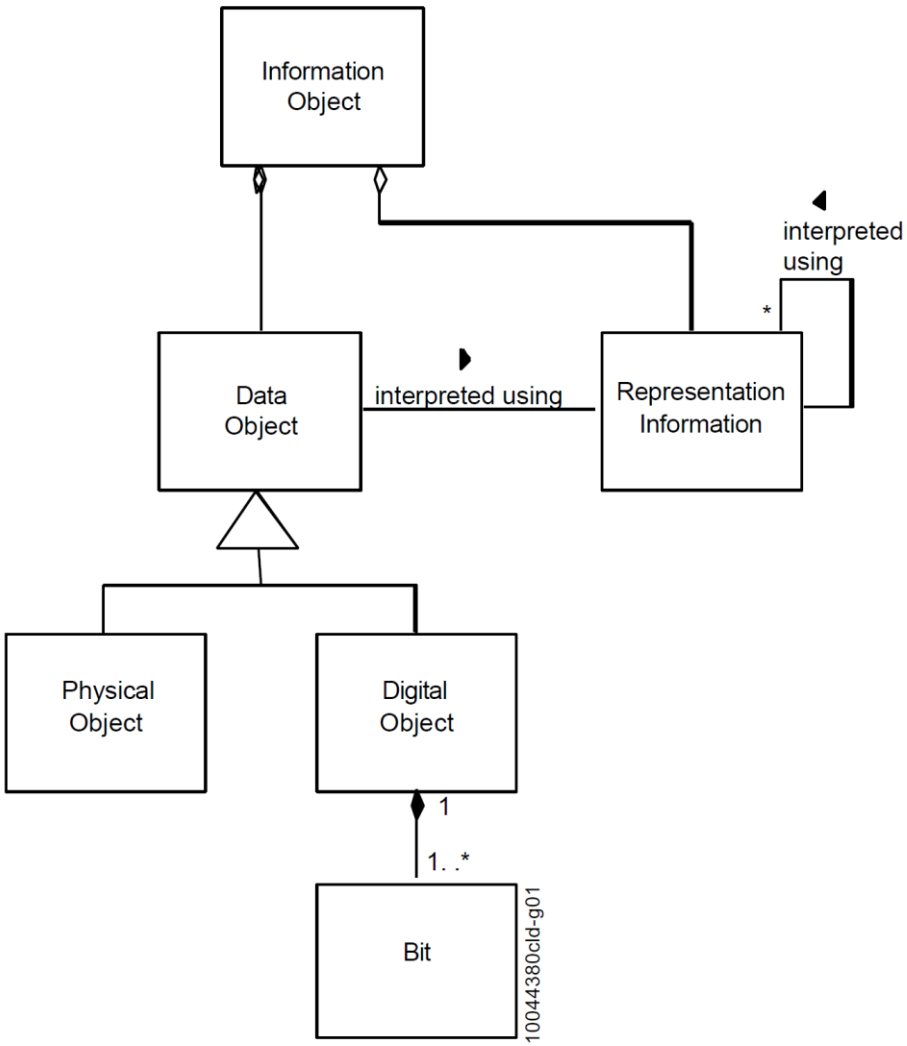
Vnímání *Datových objektů* jako smysluplné informace je umožněno pomocí *Uživatelských znalostí* (*Knowledge Base*) a *Informací o reprezentaci* (*Representation Information*), která je spojena

s *Datovým objektem*. Uživatelská znalost není vlastností OAIS systému, ale měl by ji mít každý uživatel určené *Uživatelské komunity* (*Designated Community*), pro kterou jsou data archivována primárně. Např. k pochopení textu v češtině se předpokládá, že uživatel ovládá český jazyk. Pro případ, že by uživatel takovou znalost neměl, musí existovat *Informace o reprezentaci* (*Representation Information*) k datovému objektu, tedy metadata. *Informace o reprezentaci* zajistí odpovídající zobrazení, pochopení a interpretaci obsahu digitálního objektu. Jinými slovy dává archivovanému bitstreamu smysl, popisuje jak objekt interpretovat, v jakém HW a SW jej lze zobrazit, jaké je kódování dat apod. Řekne nám např., že konkrétní řetězec znaků bitstreamu je text, včetně dalšího kontextu k pochopení obsahu. Umožní nám pochopit jak data, tak doprovodná metadata.

Informace o reprezentaci popisují samotný archivní objekt, ale mohou popisovat i metadata samotná (což je také digitální objekt sám o sobě). *Informace o reprezentaci* se liší od *Popisné informace* (*Description Information*), která musí být vytvářena ručně, kdežto *Informace o reprezentaci* je nejčastěji vytvářena automaticky pomocí různých nástrojů. *Informace o reprezentaci* může obsahovat dva typy informace – strukturální a sémantickou. Strukturální interpretuje uložené bity do typů, skupin apod. Sémantická popisuje jejich další smysl [OCLC/RLG WORKING GROUP ON PRESERVATION METADATA, 2001, s. 13]. Strukturální Informace o reprezentaci tak může být údaj o datovém formátu (např. PDF), o HW a SW nutném pro zobrazení digitálního objektu aj. Sémantická informace může popisovat, že PDF je dokument v češtině a jde o vědecký článek.

Informace o reprezentaci v reálném světě metadat je vyjádřena technickými, administrativními a ochrannými metadaty. Zajímavostí je, že dle OAIS, pokud je *Informace o reprezentaci* v digitální podobě, sama potřebuje další *Informaci o reprezentaci*, tj. o sobě samé. Tím se model OAIS dostává do nekonečného kruhu. Je potřeba si určit, kdy je informace dostatečná pro dlouhodobé uložení a pochopení obsahu v budoucnu. Protože *Informace o reprezentaci* může být užitečná pro více podobných případů (určuje např. HW a SW potřebný pro zobrazení konkrétního datového formátů a další závislosti), objevily se velmi brzy úvahy o vytvoření databáze *Informací o reprezentaci*. Jedna taková báze vznikla v projektu CASPAR, další v organizaci DCC. Za určitý druh obdobné báze lze považovat i formátové registry, jako např. PRONOM a UDFR.

Datový objekt s *Informací o reprezentaci* tvoří tzv. *Informační objekt* (*Information Object*).



Obrázek 9 – Informační objekt podle OAIS [LUPOVICI a MASANÈS, 2000, s. 5].

OAIS rozlišuje čtyři typy *Informačních objektů*:

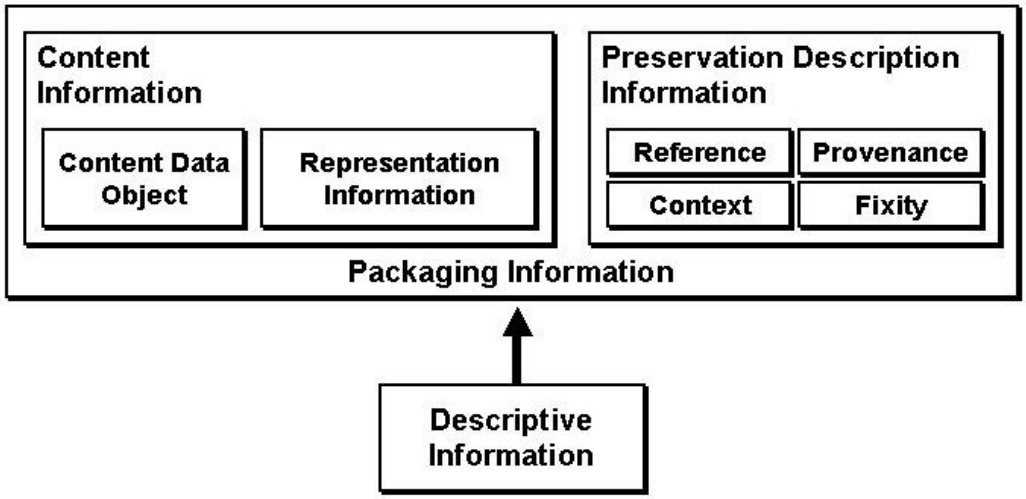
- *Informaci o obsahu* (*Content Information*);
- *Popisnou informaci pro ochranu* (*Preservation Description Information*);
- *Informaci o zabalení* (*Packaging Information*);
- *Popisnou informaci* (*Descriptive information*).

Jednotlivé typy *Informačních objektů* v reálném světě archivních systémů, dat a metadat v nich uložených obsahují – dle [OCLC/RLG WORKING GROUP ON PRESERVATION METADATA, 2001, s. 14]:

- *Informace o obsahu* (*Content Information*) – obsahuje digitální objekt (*Content Data Object*), který se archivuje a příslušnou informaci o něm, tzv. *Informaci o reprezentaci*, která napomáhá odpovídajícímu zobrazení, pochopení a interpretaci digitálního objektu. V případě *Informace o reprezentaci* tedy může jít o metadata obsahující údaje o HW a SW prostředí nutném k zobrazení digitálního objektu, o formátu digitálního objektu, jeho velikosti v kB, omezení přístupu, funkcionalitě apod.

- *Popisná informace pro ochranu* (*Preservation Description Information*) – obsahuje informace potřebné pro dlouhodobou ochranu relevantních archivních dat; OAIS rozlišuje čtyři podskupiny *Popisné informace pro ochranu*:
    - *Informace o identifikátorech* (*Reference Information*) – obsahuje identifikátory náležející k archivnímu objektu, např. ISBN, ISSN, URN:NBN, DOI aj. externí nebo vnitřní identifikátory.
    - *Informace o původu* (*Provenance Information*) – obsahuje údaje o původu archivního objektu, jeho změnách a opatřeních na něm prováděných během jeho životního cyklu, včetně informací o tom kdo, proč a kdy uvedené změny prováděl.
    - *Kontextovou informaci* (*Context Information*) – obsahuje údaje o kontextu archivovaného objektu, např. proč vznikl, vztahy k dalším objektům v archivu i mimo něj (např. typ vztahu „Is Part Of“ nebo „Has Part“ apod.).
    - *Informace o celistvosti* (*Fixity Information*) – dokumentuje procesy zajišťující neměnnost archivního objektu (nebyl nijak neoprávněně změněn).
- *Informace o zabalení* (*Packaging Information*) – spojuje archivní digitální objekt a příslušná metadata do identifikovatelné jednotky nebo balíčku.
- *Popisná informace* (*Descriptive information*) – umožňuje přístup k obsahové informaci pomocí vyhledávání; slouží jako prostředek pro použití nástrojů vyhledávání; je většinou vytvářena z obsahové informace a informace pro ochranu a vyjádřena ve schématech popisných metadat.

Výše uvedené čtyři typy *Informačních objektů* mohou tvořit *Informační balíček*, který nabývá tří podob – SIP, AIP a DIP, popsaných v kapitole 3.3.2. Celý *Informační balíček*, se všemi *Informačními objekty* včetně jejich součástí znázorňuje Obrázek 10.

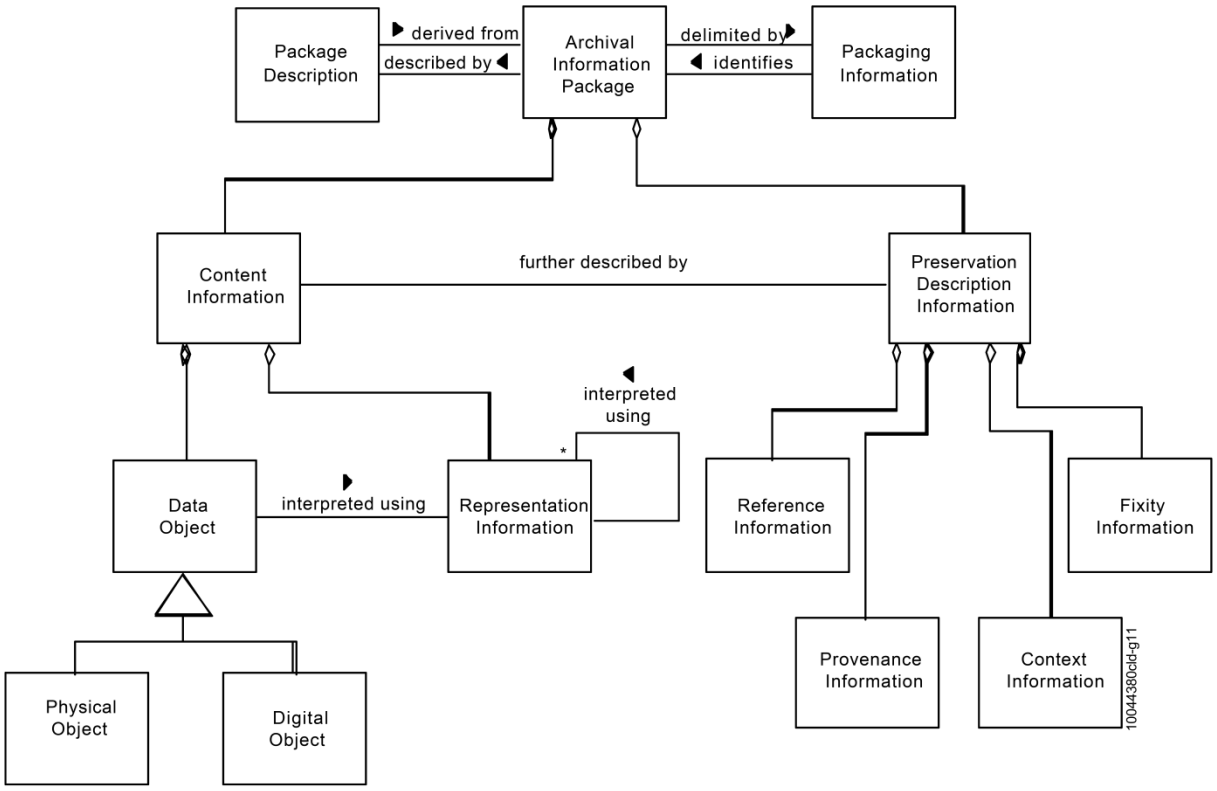


Obrázek 10 – Informační objekty tvořící Informační balíček OAIS [DAPPERT, 2009, snímek 25].

Každý ze tří balíčků obsahuje data a k nim relevantní metadata. Ta jsou ovšem v jednotlivých balíčcích různá. SIP a DIP balíčky totiž nemusejí obsahovat všechny čtyři typy *Informačních objektů*, jak jsou popsány v OAIS. Např. SIP může obsahovat pouze popisná metadata, během

vkládání do archivu vzniknou další technická metadata. AIP pak obsahuje jak popisná, technická tak administrativní metadata, která se neustále doplňují. Z pohledu OAIS tedy AIP má oproti SIP navíc např. *Popisnou informaci pro ochranu*. Obsah DIP balíčku se mění dle oprávnění uživatele. Ten může být oprávněn dostat např. pouze metadata, ale ne vlastní archivní dokument. V takovém případě v DIP balíčku chybí *Informace o obsahu*. Současný trend je takový, že již v digitalizaci vznikne co největší množina metadat, včetně technických a administrativních, která jsou dále v archivu doplněna, validována. Nejpodrobnější a nejdůležitější jsou metadata obsažená v AIP balíčku.

AIP balíček je agregací všech čtyř již zmíněných typů *Informačních objektů*, z nich každý se skládá z *Datového objektu* a *Informace o reprezentaci* – viz Obrázek 9. Tj., z logiky OAIS jsou součástí AIP: vlastní archivovaný obsah a k němu informace o něm (*Informace o reprezentaci*), dále *Popisná informace* a případná *Informace o reprezentaci* o ní samé, *Informace o zabalení* a případná *Informace o reprezentaci* a konečně *Popisná informace pro ochranu* a opět její *Informace o reprezentaci*. V reálném nasazení repozitářů odpovídajících OAIS je využívána a ukládána pouze *Informace o reprezentaci* náležející k vlastnímu archivovanému digitálnímu objektu a další metadata (ochranná, balící a popisná) [LUPOVICI a MASANÈS, 2000, s. 6]. Ne ovšem k nim náležející *Informace o reprezentaci* – viz Obrázek 10. Všechny tyto typy informací, ze kterých se skládá AIP balíček, mají svůj odraz v různých schématech ochranných, popisných, administrativních nebo technických metadat. Informační model OAIS dává základ, resp. výchozí bod, pro jejich rozvoj a dostatečnou obecnost pro použití na jakýchkoliv typech digitálních objektů.



Obrázek 11 – Detailní znázornění AIP balíčku [LUPOVICI a MASANÈS, 2000, s. 9].

4.10 Role metadat v digitálních repozitářích a LTP systémech – shrnutí

Z předchozích částí vyplývá několik oblastí a aktivit, bez kterých se běžící repozitář nebo LTP systém neobejdou. Digitální repozitář bez metadat by byl zcela nepoužitelným skladištěm dat, se kterými by se nedalo nic dělat. Příklady procesů, ve kterých jsou metadata nezbytná, jsou vyhledávání a odpovídající popis digitálních objektů. Využití dat bez popisných metadat je těžko představitelné. V současnosti nevznikají pouze popisná metadata, ale i metadata, která charakterizují objekt z více hledisek, technického, administrativního, z pohledu práv, které se k objektu vážou apod. Tyto informace jsou klíčové pro procesy kontroly integrity objektu, jeho autenticity během celého životního cyklu i jako porovnání vlastností objektu po jeho změnách (migracích apod.). Typickou aktivitou založenou výlučně na metadatech je plánování a provádění ochrany digitálních objektů s ohledem na rychlý proces stárnutí formátů a změn technologií, HW i SW pro jejich zpřístupnění. Právě na základě technických a administrativních metadat LTP systém je schopen posoudit rizika objektu hrozící, připravit a provést ochrannou aktivitu.
Správa digitálních objektů také využívá popisných, administrativních, technických metadat. Na základně těchto údajů lze objekty kontrolovat, manipulovat s nimi, vytvářet sbírky a množiny objektů určených k dalším aktivitám.

5. Implementace a vývoj používání metadat v NK ČR

5.1 Vývoj využívání metadat v NK ČR – úvod

V předchozích částech textu jsme přijali za svou myšlenku, že za metadata lze považovat katalogizační záznam. Záznamy katalogizačního rázu jsou v NK ČR vytvářeny od samého počátku její existence. V knihovnách, které NK ČR organizačně předcházely, vznikaly různé seznamy knih i s jejich popisy apod. Elektronická podoba katalogizačních záznamů se ve světě objevila v 60. letech 20. století, kdy na základě standardu MARC začaly vznikat první elektronické databáze bibliografických záznamů. Bývalé Československo, ani další země tzv. východního bloku, se tohoto vývoje díky politické situaci a izolaci od vyspělého světa nemohly účastnit, snad jen jej z povzdálí sledovat. K prvnímu zavádění počítačů do knihovnických procesů na území ČR docházelo až v 80. letech 20. století, kdy od roku 1983 začala vznikat Česká národní bibliografie na textovém procesoru Wang v maximálním počtu 10 tisíc záznamů za rok. Tomuto procesu a jeho tištěným výstupům výrazně pomohlo zavedení popisu podle pravidel ISBD. V roce 1985 se v pětiletém projektu začal využívat volně dostupný databázový systém CDS/ISIS (*Computerized Documentation Service/Integrated Set of Information Systems*) a jeho nadstavba MAKS (*Modulární automatizovaný knihovnický systém*) s novým standardem CSMARC, který odpovídal mezinárodním standardům [STOKLASOVÁ a SVOBODA, 1991]. Systém CDS/ISIS byl určený pro rozvojové země a podporovaný i distribuovaný organizací UNESCO[174]. Projekt využití CDS/ISIS a MAKS v NK ČR měl zkratku ASZF (*Automatizovaný systém zpracování fondů*) [KNOLL, 2010a, s. 20]. Od konce 80. let minulého století tedy probíhala automatizace, převod papírových katalogů na katalogy elektronické v systému CDS-ISIS. Je potřeba si uvědomit, že až do nástupu automatizovaných knihovních systémů se v české knihovnické praxi používaly ve většině národní standardy, které nereflektovaly vývoj v západních zemích, např. československý výměnný formát.[175] Nástup počítačů a samozřejmě změna politické situace v zemi tento zakonzervovaný stav začaly pomalu měnit. Velkou proměnu pro NK ČR a knihovny v ČSSR obecně znamenal přechod na MARC standard. Po přechodu na UNIMARC v roce 1994 začaly velmi rychle vznikat elektronické záznamy, elektronické katalogy lokální a později i souborné a kooperativní. NK ČR v obou případech byla hlavní metodickou institucí a podporou při implementaci těchto formátů i jejich dalším vývoji. Tyto formáty znamenaly mj. i standardizaci procesu katalogizace a možnost spolupráce na národní i mezinárodní úrovni. Zároveň podpora těchto formátů byla nutná u všech automatizovaných knihovních systémů [KAŠPAROVÁ a PSOHLAVEC, 2008, s. 16], které se u nás začaly objevovat. Poprvé také docházelo k větší a koordinované výměně digitálních záznamů, např. v rámci aktivit okolo Souborného katalogu ČR, který vznikal (resp. jeho koncepce) v letech 1993-1995 v projektu CASLIN (*Czech and Slovak Library Information Network*). Odtud byl již jen krok k mezinárodní spolupráci, která byla možná i díky tomu, že ČR přebrala standard UNIMARC jen s malými odchylkami a zároveň začala používat pravidla popisu AACR2R). ČR začala v roce

[174] Systém CDS/ISIS UNESCO vyvíjí a v mnoha jazykových mutacích nabízí dodnes. Jeho vývoj začal na konci 60. let minulého století.

[175] Např. záznamy exportované z CDS/ISIS odpovídaly normě ISO 2709 pro přenos dat, ale musely se do formátu UNIMARC převádět, což systém prováděl v přídavném převodníku [PAVLICOVÁ, 2001].

1994 jako první východoevropská země posílat své záznamy do světového katalogu WorldCat. V roce 2004 proběhl v ČR přechod z UNIMARCu na MARC21, který má větší mezinárodní podporu a interoperabilitu.

V oblasti historických dokumentů byl vývoj odlišný. Přejímání světových standardů pro zpracování knihovních jednotek nebylo tak rychlé jako v případě novodobých dokumentů, elektronický katalog historických dokumentů v dnešní podobě začal vznikat až po roce 2000. Katalogizace starých tisků si žádá více individuální přístup k jednotlivým dokumentům. Je důležité si uvědomit, že bibliografický popis tištěné knihy se koncepčně liší od popisu rukopisu, který je jednotlivinou a má jiné a větší nároky na popis. Často nelze aplikovat stejná pravidla a zásady popisu jako pro knihy moderní tištěné. Aktivity směřující k elektronickému záznamu rukopisů se čas od času objevily v Evropě nebo ve světě, např. italský program *Manus* nebo aktivity rakouské národní knihovny a bavorské státní knihovny na konci 90. let při zpřístupnění záznamů rukopisů a starých tisků [UHLÍŘ, 1999b, s. 110]. Katalogizace historických dokumentů probíhala od počátku v prostředí NK ČR v nativním standardu projektu a digitální knihovny Manuscriptorium. V rámci projektu vznikaly nové standardy metadat. Ke zpracování, katalogizaci rukopisů, starých tisků i prvotisků docházelo právě ve spojení s digitalizací. Tento přístup umožňuje provádět podrobnější popis, je však náročnější a dochází tím k tomu, že vytvořené záznamy nejsou ve stejném standardu jako záznamy ostatní z běžného katalogu NK ČR a ani nejsou jeho součástí. Jsou naopak součástí otevřeného katalogu *Manuscriptoria*[176], což se v současnosti ukazuje jako problém. Velmi významnou pro rozvoj bibliografických metadat pro historické dokumenty (zejména rukopisy), byla v tomto směru účast NK ČR v projektu MASTER (*Manuscript Access through Standards for Electronic Records*) na konci 90. let 20. století. Z hlediska typu metadat se stále jednalo čistě o metadata popisná v podobném rozsahu, jak jej známe dnes z katalogizačních záznamů.

Využívání metadatových standardů a metadatový popis digitálních objektů jak je chápeme dnes, a které jsou náplní této práce, je v NK ČR nedílně spojen se dvěma projekty, a to *Memoria (Manuscriptorium)* a *Kramerius*. *Memoria* vyrostla na prvotních digitalizačních aktivitách NK ČR v 90. letech. Byla to iniciativa vzniklá v souvislosti s provozem programu *Memoriae Mundi Series Bohemica*. Cílem bylo vytvořit virtuální badatelské prostředí pro zpřístupnění historických fondů [PSOHLAVEC, 2004, s. 40], výstupem byl systém digitální knihovny a otevřeného sdíleného katalogu historických fondů pod názvem Memoria. Konkrétní aplikace systému Memoria v NK ČR se nazývá Manuscriptorium [KNOLL, POLIŠENSKÝ a UHLÍŘ, 2004, s. 3]. Program *Memoria* byl spojen s prvními aktivitami na poli metadat, na něž později navazovaly aktivity podporující projekt *Kramerius* a typy dokumentů, kterým se věnoval. Projekt *Kramerius* vznikl na sklonku 90. let z potřeby zachránit písemné kulturní dědictví ohrožené degradací papíru, tj. vytištěné na kyselém papíře. Digitalizuje novodobé dokumenty z 19. a 20. století, převážně monografie a seriály (noviny, časopisy apod.).
Metadata pro digitální objekty začala vznikat až s prvními kroky v digitalizaci, ke kterým v NK ČR došlo v roce 1992, kdy NK ČR přistoupila k programu UNESCO *Paměť světa* (*Memory of the World/Memoriae Mundi*). Z důvodů propagace požádalo UNESCO určité instituce s vzácnými fondy o vytvoření CD-ROM s digitální kopií konkrétního díla. NK ČR výzvu přijala a úsilí dovršila

[176] V digitalizaci vznikala metadata popisná, která byla součástí digitálního objektu. Jejich popisná část byla pak také součástí Manuscriptoria ve smyslu katalogu historických dokumentů.

jako první z oslovených institucí, kdy v dubnu 1993 představila na jednání UNESCO v Paříži první kompletní digitalizovaný rukopis na CD-ROM disku na světě vůbec [KNOLL, 2009, s. 1]. V NK ČR v roce 1995 vznikla česká odnož *Paměti světa*, tzv. *Memoriae Mundi Series Bohemica* a byla vyhlášena jako národní program digitalizace historických fondů pod názvem *Národní program digitálního zpřístupnění vzácných dokumentů Memoriae Mundi Series Bohemica*. Toho samého roku vznikly první výstupy české verze *Paměti světa*, několik rukopisů prezentovaných na CD-ROM s proprietárním systémem zpřístupnění dostupným z CD, na kterém byla uložena i vlastní data.
Digitalizace jako historických dokumentů rutinní proces probíhala od roku 1996 [KNOLL, 1998], kdy bylo vybaveno pracoviště digitalizace v NK ČR. V ostatním světě byla digitalizace kulturních fondů stále také v začátcích, metadata, na která se knihovny a kulturní instituce soustředily, byla pouze popisná, i když někdy šlo o popisy velmi rozsáhlé. Z uvedeného vyplývá, že první skupinou dokumentů, které se digitalizovaly, byly ty nejvzácnější rukopisy, později i staré tisky a historické mapy. Tento profil dodnes platí pro digitální knihovnu Manuscriptorium.
Novodobá periodika a monografie z 19. a 20. století se začala digitalizovat až na sklonku 90. let 20. století (pozdější projekt *Kramerius*). Napřed šlo o periodika, později i o monografie přístupné přes digitální knihovnu Kramerius. V současné době nabízí přístup k zhruba devíti milionům stranám dokumentů. Metadatový popis i standardy metadat jsou odlišné od historických dokumentů.
V roce 2000 se programy *Memoria* a *Kramerius* staly národními programy podporovanými Ministerstvem kultury ČR. Díky tomu jsou zařazeny v programech VISK. V této souvislosti docházelo k paradoxní situaci, že ačkoliv *Manuscriptorium* i *Kramerius* byly a jsou jednou z hlavních a prestižních činností NK ČR, nebyly nikdy pevnou součástí jejího rozpočtu. Muselo se vždy žádat o finance na následující rok v rámci programů VISK6 a VISK7, případně VISK4. Ke změně došlo až v roce 2010, od kterého Ministerstvo kultury ČR poukazuje peníze přímo do rozpočtu NK ČR a ta nemusí žádat o financování těchto aktivit v rámci VISK.

NK ČR začala s digitalizací v porovnání s ostatními srovnatelnými knihovnami poměrně brzo, díky tomu prošla určitými technologickými změnami, jako byl přechod ze SGML na XML v obou zmíněných projektech. Vývoje metadatového popisu se také podstatně dotkla změna v nazírání na samotné digitalizované dokumenty z hlediska jejich dlouhodobé archivace, která se stala aktuálním tématem až s příchodem nového tisíciletí, předchozí vývoj akcentoval pouze popisná metadata.[177]
Ve využívání metadat v NK ČR lze rozlišit dvě období. V prvním období docházelo k vývoji vlastních standardů (DOBM, DTD monografie a periodika, MASTER+), všeobecně používané a standardní formáty nebyly tehdy příliš rozšířené. V období druhém je již patrná snaha o přebírání již hotových specifikací formátů a jejich případnou úpravu pro vlastní potřeby (TEI P4 a P5, MODS, PREMIS aj.).
Pro NK ČR a tvorbu metadat je typické, že vznikala stejná metadata pro archivní i pro uživatelské kopie. Nerozlišovala se odlišná podstata těchto kopií, což ale vzhledem k tomu, že ve valné většině vznikala metadata popisná, nevadilo. Později po roce 2008 vznikala shodná technická metadata ve VISK7 jak pro archivní, tak pro uživatelské kopie. V projektu NDK se počítá s tím, že metadata archivních kopií budou podstatně podrobnější, než metadata uživatelských kopií, které nebudou uložené v LTP systému.

[177] V NK ČR se nikdy cíleně nevyužívala metadata uložená uvnitř obrazových nebo jiných souborů, např. v jejich hlavičkách.

Pro obě období byly pro vývoj metadat určující zvláště výzkumné záměry, které NK ČR měla. Jde především o záměr *Digitální knihovna – produkce, ochrana a zpřístupnění digitálních dokumentů* (1999-2003)[178], ve kterém vznikla mj. současná DTD pro periodika a monografie. Dalším důležitým výzkumným záměrem bylo *Vytvoření virtuálního badatelského prostředí pro zpřístupnění a ochranu digitálních dokumentů* (2004-2010), ve kterém probíhaly specifikace a úpravy metadat pro historické dokumenty i pro ty novodobé.

5.2 Národní knihovna ČR – období první (1996-2003)

Období 1996-2003 ohraničuje dobu, kdy NK ČR začala s digitalizací ve větší míře v rámci své strategie a různých projektů. Pokrývá etapu vzniku vytváření prvních standardů a také přechod ze SGML na XML a s tím spojené zavedení nových standardů metadat. Tj. opuštění standardu DOBM a jeho náhradu za DTD periodika a monografie u novodobých dokumentů a zavedení nového standardu MASTER+ u dokumentů historických od roku 2003.

V rámci digitalizace nevznikala jen data a popisná metadata, ale jednoduchý nebo komplexní digitální objekt (intelektuální entita), který tvoří kompletní náhradu (kopii) fyzického dokumentu. Komplexní digitální objekt[179] je tvořen třemi úrovněmi, které lze rozlišit nejen u historických dokumentů, ale i u novodobých dokumentů v digitálních knihovnách nebo archivech.

1. První úrovní jsou samotná data – tj. soubory obrazů (archivní kopie, uživatelské kopie, náhledy v různých formátech), soubory plných textů (OCR nebo TEI soubory v různých formátech), případně další doprovodné objekty (např. DTD soubory, ICC profily v komplexním digitálním dokumentu Manuscriptoria).
2. Druhou úrovní je popis komplexního digitálního objektu. Tady se nejedná pouze o klasická bibliografická metadata dokumentu, ale i o popis jednotlivých logických částí (kapitol, stránek apod.); o vyjádření logické i fyzické struktury, tedy z dnešního pohledu strukturální metadata, která byla považována za popisná a vyprofilovala se jako samostatná skupina později. Do této části by patřila i administrativní metadata včetně technických, tato ovšem nebyla brána v úvahu jako samostatný typ metadat (až později od roku 2002). V procesu digitalizace a přípravy dokumentů pro digitální knihovnu Manuscriptorium byly využívány pro bibliografický popis, popis struktury a částečný technický popis standardy DOBM, MASTER, MASTER+, METS a TEI P5. V digitalizaci novodobých dokumentů se používal standard DOBM a poté DTD periodika a DTD monografie.

5.2.1 Metadatový popis historických dokumentů

NK ČR byla průkopníkem na poli digitalizace a popisu starých tisků a rukopisů v České republice i v Evropě. Postupy i standardy popisů z NK ČR se staly základem i pro jiné knihovny, např. Univerzitní knihovna Graz [KNOLL, 1998]. Je nutno uznat, že jistým specifikem digitální knihovny historických dokumentů, je obsáhlost popisů, které lze k jednotlivým historickým dokumentům vytvořit. Manuscriptorium, které se často prezentuje spíše jako virtuální badatelské prostředí než jako digitální knihovna, se snaží vyjít vstříc odborné badatelské obci, která má vysoké nároky na

[178] http://www.isvav.cz/h10/researchPlanDetail.do?rowId=MK0CEZ99F2001
[179] Skládá se z dalších digitálních objektů.

práci s historickými texty. Je tu jasná orientace na uživatele, kterému je předkládáno prostředí a dokumenty zpracované mnohdy se souvislostmi a dodatečnými sekundárními zdroji tak, že se blíží elektronické edici. Důraz na bibliografický popis dokumentů je i důsledkem toho, že digitální knihovna Manuscriptorium funguje jako katalog rukopisů a starých tisků.

Z počátku digitalizace se zdálo, že hlavním typem metadat, která by se měla vytvářet pro zdigitalizované dokumenty, jsou metadata popisná. Ta popisují intelektuální entitu (knihu, svazek, apod.) a napomáhají vyhledání i prezentaci konkrétního dokumentu v digitální knihovně. Jde o obdobu katalogizačního záznamu. Brzy se však ukázalo, že je nutno vytvářet a ukládat i údaje přesahující popisná metadata. Konkrétně údaje o struktuře, které napomáhají prezentaci dokumentu v jeho skutečné struktuře odpovídající předloze. Struktura tvoří kostru dokumentu a jeho jednotlivých částí. Rodící se Internet a jeho technologie nabízely možnost využití odkazování dat z popisné části dokumentu. Nabízející se HTML ovšem dokáže vyjádřit pouze formální vlastnosti, není vhodné pro popis obsahu dokumentu, kde je potřeba vyjádřit různé typy jednotlivých částí.

Při vývoji metadatových standardů pro popisy digitálních dokumentů v NK ČR se, i přes možnost pouze formálního popisu/markupu, začínalo s HTML. Počínaje rokem 1996 vzniklo několik verzí specifikací a využití HTML bylo dokonce doporučeno UNESCO v subkomisi pro program *Paměť světa*. HTML bylo doplněno o tzv. doplňkové tagy, které jeho možnosti rozšiřovaly. Pomocí HTML s přídavnými tagy bylo popsáno několik desítek rukopisů, převážně v roce 1996. Popis pomocí HTML se dostal do verze 2.0, již pod názvem DOBM, když v roce 1997 na doporučení komise *Paměti světa* bylo doporučeno orientovat se na SGML (*Standard Generalized Markup Language*) [KNOLL, 1997]. Díky zkušenostem s tagy v HTML již bylo relativně jednoduché stanovit jaké údaje je nutno mít v SGML. I v NK ČR došlo tedy k přechodu na SGML. Jinými slovy: „*Vznikla potřeba zkombinovat prezentační vlastnosti HTML s možnostmi popisu obsahu, struktury, technických vlastností dat atp., které dávalo pouze čisté SGML.*“ [KNOLL, 2009, s. 2]

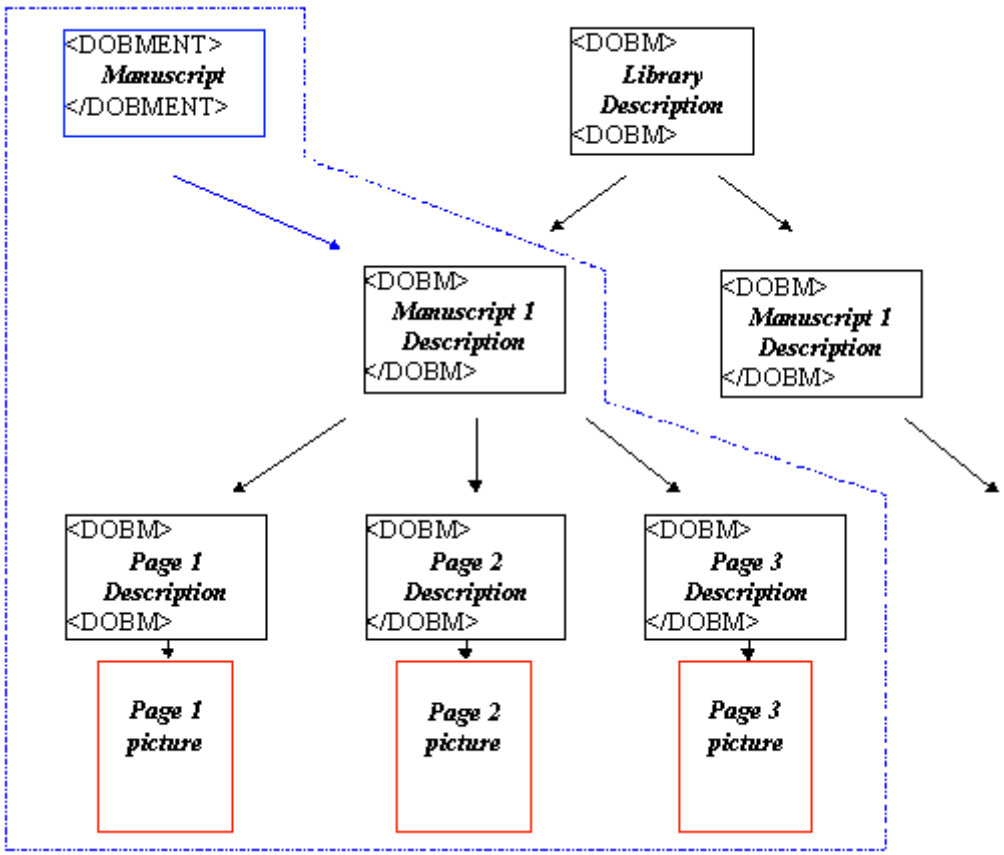
5.2.1.1 DOBM SGML

DOBM SGML (*Digitized Old Books and Manuscript*, dále jen DOBM) byl první metadatový popis starých tisků v NK ČR v syntaxi univerzálního značkovacího metajazyka SGML. Vznikl v rámci projektu *Memoria Mundi Series Bohemica*, financovaného Ministerstvem kultury ČR. Z dnešního pohledu šlo o metadatový kontejner pro „zabalení“ dat i metadat. Standard vznikal v NK ČR (Adolf Knoll) ve spolupráci s AiP Beroun v letech 1996-1998 v rámci VaV projektu *Archivace a zpřístupnění vzácných dokumentů s využitím digitální technologie*. V roce 1999 doporučila komise UNESCO *Paměť světa* metadatový standard DOBM v jeho druhé a definitivní verzi jako mezinárodní standard pro popis rukopisů v programu *Paměť světa*, což byl výrazný úspěch.[180] V této době neexistovaly dnes standardní a běžné standardy založené na XML. Standard ovšem díky promyšlené struktuře předznamenával využití XML. Záznamy ve standardu DOBM SGML byly v NK ČR tvořeny v letech 1998 až 2002.

Obecná struktura DOBM záznamu byla vyjádřena v SGML pomocí standardního DTD souboru (DOBM.DTD). Samotný DOBM standard umožňuje definovat vlastní DTD. DOBM.DTD definovalo pouze obecné elementy, bylo proto potřeba pro každý typ dokumentu (monografie, periodikum, rukopis, zvukový záznam, sbírka) definovat použití těchto obecných elementů. Toto bylo zařízeno

[180] Doporučený standard DOBM byl oficiálně prezentován na CD-ROMu, který obsahoval výklad a popis formátu, tak veškerou dokumentaci – viz [KNOLL a PSOHLAVEC, 1999].

pomocí souboru ENTER.SGM, který obsahoval vlastní DTD nazvané DOBMENT.DTD. Bylo tak možné vyjádřit strukturu každého typu dokumentu včetně popisných elementů použitých pro každou část této struktury [KNOLL, 2000, s. 3] – viz Obrázek 12. Bibliografický popis dokumentu byl uložen v souboru DESCR.HTM, který je součástí metadat a byl linkován ze souboru DOBMENT. Struktura standardu a tedy výsledného DOBM záznamu je relativně složitá, metadatový popis je složen z více DOBM záznamů. Cílem DOBM bylo vytvořit reprezentaci předlohy (knihy, rukopisu apod.). Tj., standard popisoval jak celý dokument (titul), tak jeho části (kapitoly, stránky). Na každou úroveň existoval jeden DOBM soubor – viz Obrázek 12.



Obrázek 12 – Obecné možnosti tvorby úrovní z jednotlivých DOBM souborů [KNOLL a VOMLEL, 1999b].

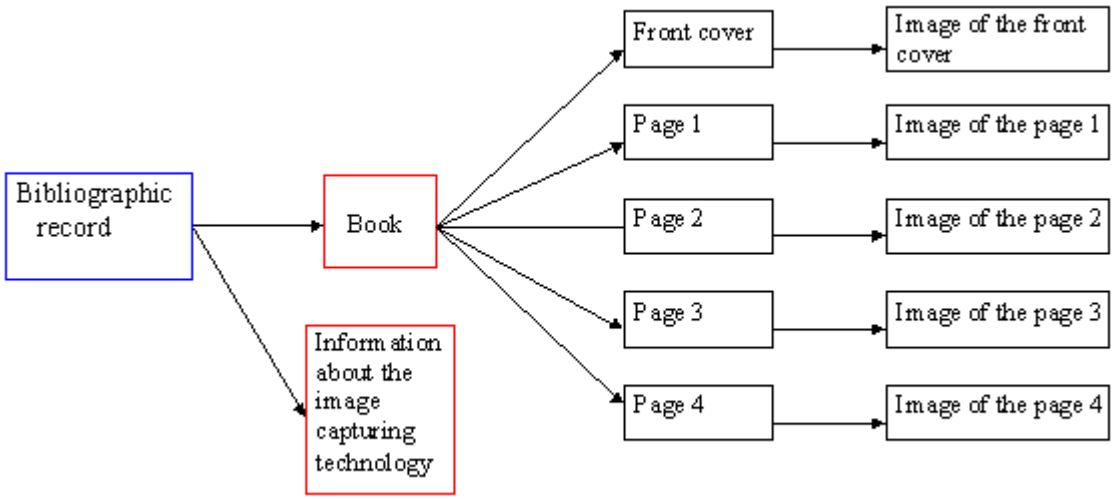
DOBM SGML standard byl postaven na HTML verzi 2 doplněné o další elementy (rozšíření HTML o obsahově orientovaný markup). Oproti HTML, které umožňuje popsat pouze vizuální formátování dokumentu, bylo potřeba vyjádřit i obsah dokumentu, kategorie důležitosti, kategorie jednotlivých částí (DOBM úrovní) a typy jednotlivých částí dokumentu, které tvoří určitou hierarchii. Klasifikace metadat z hlediska důležitosti byla umožněna díky doplňujícím elementům. Šlo o následující elementy – příklady částečně převzaty z [KNOLL a VOMLEL, 1999b]:

- **DOBM.DOC**
  - označoval kategorii DOBM souboru (např. PAGE, BIBLDESCR aj.)
  - např. <DOBM SPEC="NKP//MANUSCRIPT 2.1" CTGLABEL=PAGE NAME="Page" LANG=EN>
- **DOBM.DX**
  - označoval typ jakéhokoliv metadatového řetězce znaků

  - uvnitř elementu <DOBM.DX> mohly být použity tagy HTML, nebo tam mohly být vloženy další <DOBM.DX> nebo <DOBM.DATA> elementy
  - např. <DOBM.DX CTGLABEL=DATOFPUBL NAME="DateofPublication" TYPE=DATE><B>1756</B></DOBM.DX>
- **DOBM.DATA**
  - označoval typ odkazovaného datového souboru
  - např. <DOBM.DATA HREF = „high/0001r.jpg" CTGLABEL = HIGHQ NAME= "HQpicture">
- **DOBM.REFERENCE**
  - označoval pozici jednotlivých DOBM souborů ve struktuře dokumentu
  - např. <DOBM.REFERENCE HREF = „0001r.htm" CTGLABEL=PAGE NAME = "Page">

Metadatový popis rukopisů musel být schopen vyjádřit strukturu dokumentu tak, aby odpovídala struktuře předlohy a měla určitou flexibilitu pro různé možnosti vyjádření struktury. Metadata pro rukopisy tak byla rozdělena do následujících kategorií – viz Obrázek 13, každou z nich tvořil jeden DOBM záznam [KNOLL a MAYER, 1999][181]:

- **BIBLDESCR** – bibliografický popis rukopisu nebo starého tisku – obsahoval obdobné údaje jako katalogizační záznam a odkazy na část popisující technická metadata (TECHDESCR) a na soubory tvořící knihu (BOOK);
- **TECHDESCR** – technický popis – obsahoval údaje o procesu digitalizace, zařízení, na kterém se digitalizace prováděla, rozlišení apod. Pro rozlišení typů údajů se používaly DOBM.DX elementy, např. <DOBM.DX CTGLABEL="CAPTURE" NAME="Capture">Digitized by the Kodak DCS 460 camera.</DOBM.DX>;
- **BOOK** – tento soubor popisoval knihu jako celek a obsahoval údaje nezaznamenané v části BIBLDESCR; obsahoval také odkazy na všechny stránky a části dokumentu včetně jejich popisů;
- **PAGE** – tento soubor představoval jednu stránku knihy a popisoval stránku předlohy; součástí popisu byly také odkazy na vlastní digitální soubory stránky – různé kvality obrazu.



Obrázek 13 – Struktura metadatového popisu rukopisů [KNOLL a MAYER, 1999].

[181] Podobně specifikovány byly i další typy dokumentů, pro které vznikla DOBM specifikace.

Výhodou standardu DOBM byla možnost vytváření a přidávání jakýchkoliv dalších popisných objektů. DOBM obsahovalo pravidla definování objektů včetně jejich názvu, významu i vlastností. DOBM počítalo také s popisem odkazovaných vnějších datových objektů. To bylo možné využít pro popis různých jiných objektů náležejících ke komplexnímu digitálnímu objektu (např. náhledy aj.) [KNOLL, 1998]. Tento princip je dnes využíván i v moderních kontejnerových standardech metadat, jako je např. METS. Finální komplexní DOBM digitální dokument (jedna intelektuální entita, např. rukopis) je ve finální podobě komplex popisných souborů SGML DOBM a samotných dat.

DOBM měl ale také několik nevýhod. Technická metadata v DOBM byla minimální a nebyla strukturovaná, což ztěžovalo jejich využití, zvláště při nepřesnostech v zápisu hodnot těchto elementů. Technická metadata ve strukturované formě se začala používat až s přechodem na standard MASTER+ u historických dokumentů. U novodobých dokumentů byla bohužel nestrukturovanost technických údajů přenesena i do DTD monografie a periodika. Další nevýhody se ukazovaly při rutinním zpracování a používání vytvořených metadat v běžném provozu digitální knihovny. Jak uvádí Zdeněk Uhlíř: *„Princip pevně strukturovaného popisu DOBM využívajícího SGML … má sice výhodu v tom, že je jednoduchý, a usnadňuje tudíž rutinní až industriální práci. Nevýhodou však je to, že využívá pouze tvrdě strukturovaných dat na jedné nebo vůbec nestrukturovaných dat na druhé straně: bibliografické údaje a údaje o některých snadno typizovatelných vnějších znacích jsou ve formě tvrdě strukturovaných dat, zatímco ostatní údaje včetně údajů o intelektuálním obsahu originálního dokumentu jsou v podobě volného textu. V případě podrobného … popisu se tak klade překážka … sofistikovanějšímu vyhledávání.“* [UHLÍŘ, 2006, s. 5] Toto byl jeden z důvodů, proč se od pevně strukturovaného popisu ve standardu DOBM přešlo ke standardu MASTER.

Standard MASTER má pevně danou syntaxi a umožňuje využívání předepsaných elementů na různých (i více) horizontálních i vertikálních úrovních popisovaného dokumentu, např. při popisu stránky i titulu. Standard tak lépe odpovídá potřebám popisu materiálů, které jsou stejného typu (např. rukopisy), ale mohou se od sebe podstatně lišit – strukturou, rozsahem apod. – viz [UHLÍŘ, 2006, s. 5]. Je tak možné vytvořit krátký popis dokumentu, který má minimální strukturu nebo málo popisných údajů a naopak lze vytvořit i popis jdoucí do hloubky a detailně strukturovaný.

Obecný standard DOBM SGML bylo možné použít na popis jakéhokoliv typu dokumentu. Vedle popisu rukopisů a starých tisků (viz Obrázek 13) vznikly podobné konkrétní specifikace pro novodobá periodika a monografie, zvukové nahrávky a sbírky dokumentů[182]. K širšímu využití metadatových standardů pro zvukové nahrávky a sbírky dokumentů, ani později ve XML podobě, nikdy nedošlo[183]. Naopak z metadatových standardů DOBM SGML pro novodobá periodika a monografie vznikla v průběhu roku 2003 dodnes používaná DTD periodika a DTD monografie[184].

5.2.1.2 XML a standardy MASTER a MASTER+

Na přelomu tisíciletí se stále více a více prosazovalo XML. Nakonec v oblasti metadat převládlo, mj. i díky tomu, že dosavadní standardy, jako např. HTML, nedokázaly zaznamenat typ obsahu ani

[182] http://digit.nkp.cz/techstandards_cz.html

[183] Byl použit pouze pro testovací popis zvukových magnetofonových pásek zaznamenávajících události studentského listopadu 1989.

[184] http://digit.nkp.cz/techstandards_cz.html

strukturu metadat, a SGML bylo příliš komplikované a komplexní. Metadata ve formátu XML se dají lehce strukturovat a lehce technicky zpracovávat. XML zcela převládlo v oblasti vyjádření metadat nejen v paměťových institucích – více viz kapitola 4.6. V době vzniku standardu DOBM ještě XML neexistovalo.

Již na přelomu let 2001/2002 bylo jasné, že alternativou k DOBM, které bylo podmnožinou SGML, se stále více stává XML (taktéž podmnožina SGML). I vzhledem k rychlému rozšíření a podpoře prohlížečů i vzniku nástrojů pro tvorbu a zpracování XML, bylo v NK ČR rozhodnuto o nasazení jazyka XML a o konverzi DOBM do XML podoby [KNOLL a PSOHLAVEC, 2002, s. 2]. XML se ukazovalo jako velmi vhodné pro vyjádření struktur digitálních dokumentů i pro vyjádření bibliografických nebo technických metadat a jejich výměnu mezi jednotlivými systémy.

Vedle všeobecného akceptování XML byly hlavním důvodem přechodu potřeba návaznosti a využití aktuálních standardů, jmenovitě výstupů z projektu MASTER (*Manuscript Access through Standards for Electronic Records*) a DIEPER (*DIgitised European PERiodicals*)[185] a částečná snaha o vytváření podrobnějších technických metadat. Evropského projektu MASTER se NK ČR účastnila v letech 1999-2001 jako jeden ze šesti řádných partnerů. Cílem bylo vytvořit standard pro elektronickou katalogizaci rukopisů a vybudovat prototyp elektronického katalogu rukopisů včetně ověření funkčnosti standardu za použití různých softwarových nástrojů [RÁKOCY, 2008, s. 13]. Zajímavostí je, že standard MASTER byl původně vytvářen na platformě SGML, až v průběhu projektu bylo rozhodnuto o využití XML. Výstup projektu byl formát na obou platformách. Více o projektu a o návazném standardu MASTER viz [UHLÍŘ, 2002a]. V NK ČR došlo k testování finálního MASTER standardu určeného k bibliografickému popisu (MASTER.DTD vycházející z TEI P4) v rámci VaV *Optimalizace archivace a zpřístupnění digitálních dat* [KNOLL a PSOHLAVEC, 2002, s. 8]. Ostré nasazení standardu proběhlo roku 2003 v otevřeném katalogu historických dokumentů nově vzniklého systému Manuscriptorium.[186] V první fázi Manuscriptorium obsahovalo pouze otevřený katalog historických dokumentů, později, po vzniku standardu MASTER+, začalo fungovat jako digitální knihovna nejen rukopisů, ale i starých tisků, prvotisků a historických map.

V projektu MASTER se původně počítalo pouze s návrhem metadatového záznamu pro bibliografický popis, ovšem řešitelé se dotkli i problematiky propojení bibliografického záznamu se samotnými daty [UHLÍŘ, 2007], které by tvořilo kompletní dokument. Na základě těchto snah byl vytvořen standard MASTER+, jako rozšíření standardu MASTER. MASTER+ byl určen pro spojení katalogového záznamu (tj. popisných metadat) a datových objektů (obrazové soubory v různé kvalitě) – viz Obrázek 15. Datové objekty mohou být uloženy i přímo v repozitářích přispěvatelů do Manuscriptoria. V programu *Memoriae Mundi Series Bohemica* se MASTER+ využíval pro bibliografický i strukturální popis historických dokumentů. Tj. pro vytvoření virtuálních digitálních kopií dokumentů v digitální knihovně Manuscriptorium a pro digitální objekty uložené na CD-ROM nebo datovém úložišti. Záznamy, resp. digitální dokumenty ve standardu MASTER+ byly v NK ČR vytvářeny od roku 2002. MASTER+ byl záhy po svém vzniku vybrán jako standard pro národní program *Memoriae Mundi Series Bohemica* (VISK6).

Podoba nových standardů byla plánována a implementována tak, aby byla zpětně kompatibilní se standardem DOBM. Přechod znamenal vytvoření DTD nového standardu MASTER+ (MSNKAIP.DTD/MSNKAIP.XSD) a postupné úpravy stávajících nebo vývoj nových nástrojů na

[185] http://cordis.europa.eu/libraries/en/projects/dieper.html

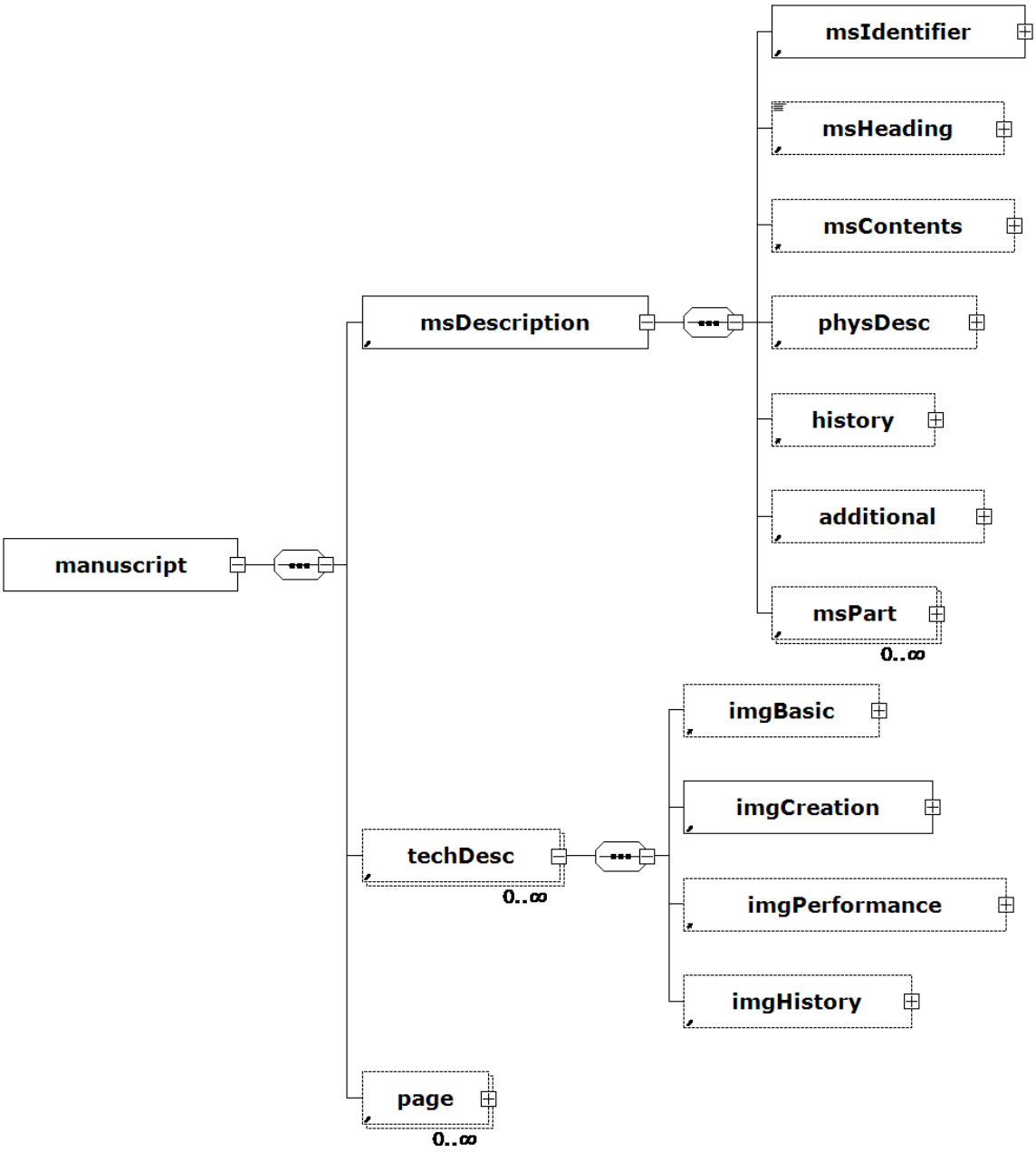
[186] Aktuální verze formátu je k dispozici tady http://digit.nkp.cz/techstandards_cz.html.

vytváření a zpřístupnění záznamů v novém standardu [KNOLL a PSOHLAVEC, 2002, s. 9]. Během roku 2003 došlo k převodu výroby celého digitálního dokumentu na platformu XML. Ve stejném roce také došlo k migraci všech metadatových záznamů starých tisků a rukopisů ze SGML na tuto platformu. Přechod na XML u novodobých dokumentů i u historických dokumentů byl spojen i s přechodem na nové způsoby zpřístupnění, tj. vytvoření samostatných aplikací Kramerius a Manuscriptorium v roce 2003 a 2004.

Standard MASTER+ sestával ze tří elementů vrchní úrovně:

- <msDescription> – bibliografický popis ve standardu MASTER z nějž byla použita jen jeho část tvořená elementem <msDescription>;
- <techDesc> – technická metadata vycházející z draftu NISO *Technical Metadata for Digital Still Images* a popis nastavení kamery/skeneru ve specifikaci DIG35 – viz [COVER, 2002];
- <page> – popisné údaje o jednotlivých stranách rukopisu, tato část vychází nejvíce z původního DOBM.

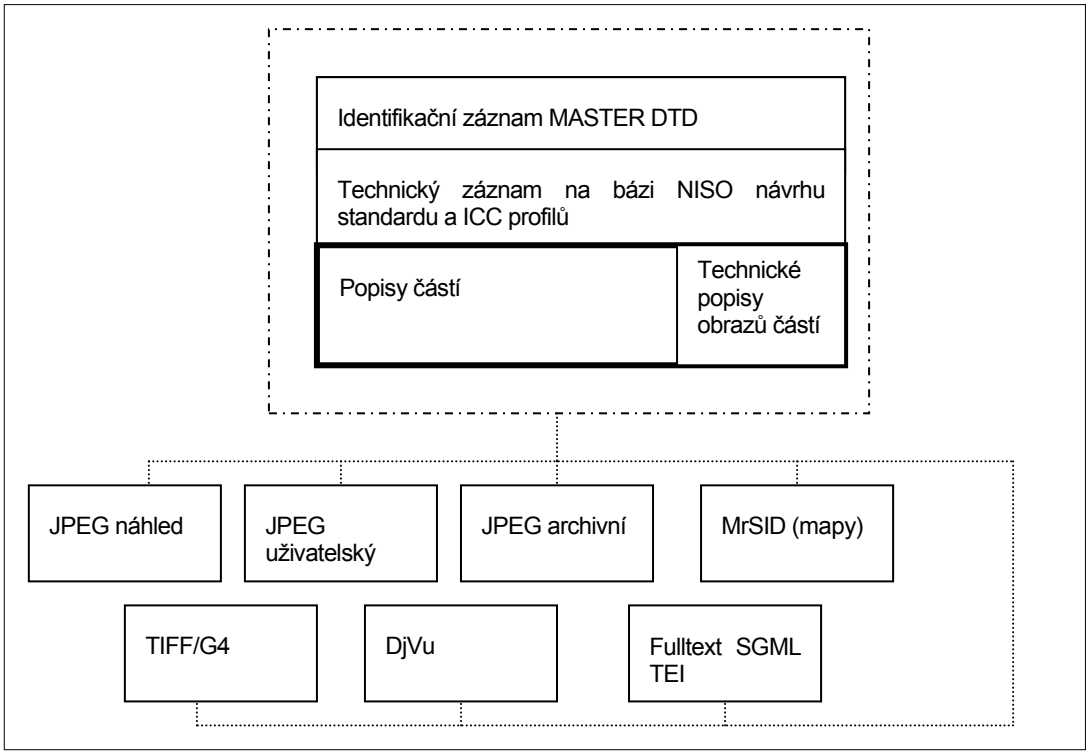
Podrobnější pohled na strukturu jednotlivých elementů vrchní úrovně MASTER+ viz Obrázek 14.



Obrázek 14 – Struktura elementů standardu MASTER+ (kontejnerového standardu).

DTD standardu MASTER+ (MSNKAIP.DTD) se skládala z více různých DTD – porovnej s poznámkou 1 v [KNOLL, 2003, s. 239] a Obrázkem 15:

- MASTERX.DTD – bibliografický popis rukopisu ve standardu MASTER XML TEI;
- DIG35.DTD – technická metadata;
- MSNKAIP_DEVICE.DTD – technický popis digitalizace rukopisu;
- MSNKAIP_PAGE.DTD – popis strany rukopisu.



Obrázek 15 – Struktura metadatového záznamu rukopisu ve standardu MASTER+ [KNOLL a PSOHLAVEC, 2002, s. 8].

Z Obrázku 15 lze vidět, že součástí popisu jsou také podrobnější technická metadata. Je to důsledek toho, že zároveň se změnou standardu bibliografických metadat a struktury celého dokumentu, se přistoupilo k dalšímu podstatnému doplnění metadat, a to o metadata technická ve strukturované formě. Bylo rozhodnuto využít pokud možno uznávaný standard. Volba padla tehdy ještě na návrh standardu datového slovníku NISO *Technical Metadata for Digital Still Images* ve verzi z roku 2001[187]. Tento standard je velmi rozsáhlý a dokáže postihnout z technického hlediska jak proces výroby digitálního obrazu, tak nastavení skenerů/kamer, nastavení profilů atp. V rámci implementace bylo také využito a lehce doplněno DTD podle specifikace DIG35 (*Metadata for Digital Images*) publikované poprvé v červnu 2000 – verze ze srpna 2000 viz [DIGITAL IMAGING GROUP, 2000]. Specifikace DIG35[188] byla použita na popis nastavení kamery. Elementy pocházející z DIG35.DTD byly jako dceřiné elementy rodičovského elementu <CAMERA_SETTINGS> vloženy do kompletního DTD pro technická metadata, které vzniklo dle specifikace NISO. XML podoba (DTD) pro zápis jednotlivých prvků draftu NISO standardu doplněných o popis nastavení skenovacího zařízení (kamery) ve standardu DIG35 se tak stala součástí popisu rukopisu. V uvedené době neexistovala dnes běžná podoba zmíněného NISO standardu zapsaná v XML – schéma MIX[189] (*Metadata for Images in XML*). Jednalo se tak vlastně o předchůdce pozdějšího XML standardu MIX, který vznikal od roku 2001 v Kongresové knihovně v USA.

[187] Návrh standardu se stal v roce 2006 normou ANSI/NISO Z 39.87 2006.
[188] Schéma DIG35 obsahovalo pět hlavních částí (basic image; image creation; content description; history a intellectual property rights), které se velmi podobají pozdější specifikaci schématu MIX. Celkově mělo přes 150 elementů.
[189] http://www.loc.gov/standards/mix/

Standard NISO uvedený výše je zaměřen na detailní popis obrazových souborů. V programu *Memoriae Mundi Series Bohemica* bylo hlavním cílem digitalizace zpřístupnění. Proto se důležitost přikládala více procesu vzniku a procesu zobrazení. Důvodem byla snaha zachytit přesné nastavení procesů i technologií skenování tak, aby bylo možno zdigitalizované obrazy na základě těchto technických metadat zpřístupňovat v nejvěrnějším barevném podání. Součástí technického metadatového popisu dokumentů v Manuscriptoriu jsou i údaje o ICC profilech a barvách.[190] Celý potenciál standardu NISO nebyl bohužel využit, mj. i z obavy, aby nedocházelo k samoúčelnému nárůstu objemu dat. V dokumentu *DTD pro projekt Memoriae Mundi Series Bohemica* uvádějí autoři Ing. Psohlavec a Ing. Kučera následující: *„Popis MMSB je popisem konkrétního strukturovaného dokumentu, kde procesy jsou aplikovány na celou skupinu obrazů, a hodnoty jsou obvykle pro celé skupiny obrazů společné. Jednotlivé obrazy se liší jen v jemných detailech své geneze. Prvotní význam vznikajících dokumentů je v informačním obsahu neseném obrazem. Z tohoto hlediska příliš podrobné technické popisy nemají svého adresáta a redundance dat implicitně obsažených v obrazech samotných není zcela odůvodnitelná. Detailní individuální technické vlastnosti jednotlivých obrazů nejsou z hlediska hlavního předpokládaného uživatele důležité.“* [KUČERA a PSOHLAVEC, 2001, s. 3]

Technická metadata mohla být plněna na úrovni celého rukopisu, pokud byl digitalizován kompletně na stejném zařízení, nebo v případě odlišností u jednotlivých obrazů, mohl být element <techDesc> opakován na úrovni každé stránky. Je jen logické, že v případě plnění technických metadat na úrovni celého rukopisu, nemohly být plněny údaje týkající se jednotlivých obrazů, resp. popisující jejich vlastnosti, jako např. velikost souboru, kontrolní součet aj. Tyto údaje tedy plněny nebyly a vrchní úroveň technického popisu se používala pro zachycení nastavení skenovacího zařízení, údajů o digitalizující instituci, tj. údajů, které byly pro všechny obrazy v rukopisu společné. Přitom právě velikost souboru, kontrolní součet jsou pro dlouhodobou ochranu a integritu samotného souboru velmi podstatné. Z toho je jasně patrné, že technická metadata v projektu *Manuscriptorium* měla jinou úlohu, než bychom v dnešní době očekávali. Celé to shrnuje v projektové zprávě z roku 2002 Adolf Knoll: *„Nechceme, aby se šířka zaznamenávaných [technických] dat stala samoúčelnou, naopak stále je třeba mít na zřeteli prvotní účel digitalizace, kterým je služba primárním dokumentům (jejich ochrana před používáním) a badatelům (zpřístupnění informací, vznik nového informačního prostředí).“* [KNOLL a PSOHLAVEC, 2002, s. 14]

Čtyři hlavní elementy technického popisu[191] standardu MASTER+ byly následující:

- <imgBasic> – obsahuje základní údaje o obrazovém souboru (např. formát, MIMEtype, kompresi, velikost souboru aj.);
- <imgCreation> – obsahuje údaje o podmínkách digitalizace (např. nastavení kamery/skeneru, údaje o skeneru, osobě/instituci provádějící snímkování apod.);
- <imgPerformance> – obsahuje údaje o parametrech digitálního obrazu (např. barevnou hloubku, rozměry obrazu apod.);
- <imgHistory> – obsahuje informace o úpravách provedených na naskenovaných obrazech (např. kdy, kdo a na jakém SW konkrétní změnu dělal).

[190] Tento typ údajů je běžný i v digitalizaci novodobých dokumentů, jejich uvádění je ovšem vedeno snahou o dokumentaci procesu skenování a následné aktivity dlouhodobé ochrany digitálních dat.
[191] Podobnou strukturu měla i první verze formátu MIX.

Pouze následující elementy byly doporučeny k využití pro úroveň popisu celého rukopisu [KUČERA a PSOHLAVEC, 2001, s. 27 a dále]:

- MIMEtype,
- format,
- ScanningAgency,
- DeviceSource,
- Sensor,
- DateTimeCreated,
- XsamplingFrequency ,
- YsamplingFrequency,
- SamplingFrequencyUnit,
- BitsPerSample,
- SamplesPerPixel,
- Profiles,
- DateTimeProcessed,
- SourceData ,
- ProcessingAgency,
- ProcessingSoftware,
- ProcessingActions,
- dále rozpis informací o nastavení pracoviště digitalizace, jako např. čas expozice, typ osvětlení, vyvážení bílé apod.

Využití elementů se samozřejmě lišilo a vyvíjelo dále. Ne všechny byly vždy používány. Pro úroveň jednotlivých stránek nebyl doporučen žádný element, jejich použití bylo volitelné dle MNSKAIP.XSD. Z dnešního podhledu logické dlouhodobé ochrany je tento rozsah nedostatečný, zvláště pokud je pouze na vrchní úrovni a ne pro jednotlivé obrazy.

Autoři specifikace technických metadat bohužel nereflektovali, že jedním z hlavních smyslů datového slovníku NISO uvedeného výše je uchovat informace o obrazech podstatné pro dlouhodobou ochranu digitálních dat. Co nejpodrobnější technický popis by měl mít každý zdigitalizovaný obraz, i když se liší pouze nepatrně. Umožní to jejich správu, manipulaci na základě konkrétních technických vlastností, kontrolu integrity apod. Každý obraz se navíc liší v několika podstatných vlastnostech, jako jsou kontrolní součty, které v Manuscriptoriu nevznikaly až do roku 2010!

Technická metadata tedy nevznikala s cílem umožnění dlouhodobé ochrany vlastních digitálních obrazů. I přesto jde o první standardní použití technických metadat v NK ČR a rozsah elementů z NISO *Technical Metadata for Digital Still Images* a z DIG35.DTD, které byly opravdu plněny, byl velkým pokrokem.

5.2.2 Metadatový popis novodobých dokumentů

5.2.2.1 DOBM SGML a jeho využití pro novodobé dokumenty

Do roku 1999 v programu *Kramerius* probíhalo pouze mikrofilmování dokumentů ohrožených kyselostí papíru. Digitální kopie mikrofilmů pomocí hybridní metody začaly vznikat v projektu *Digitalizace mikromédií*, který NK ČR vedla v letech 1997-1999. V rámci projektu byla zakoupena

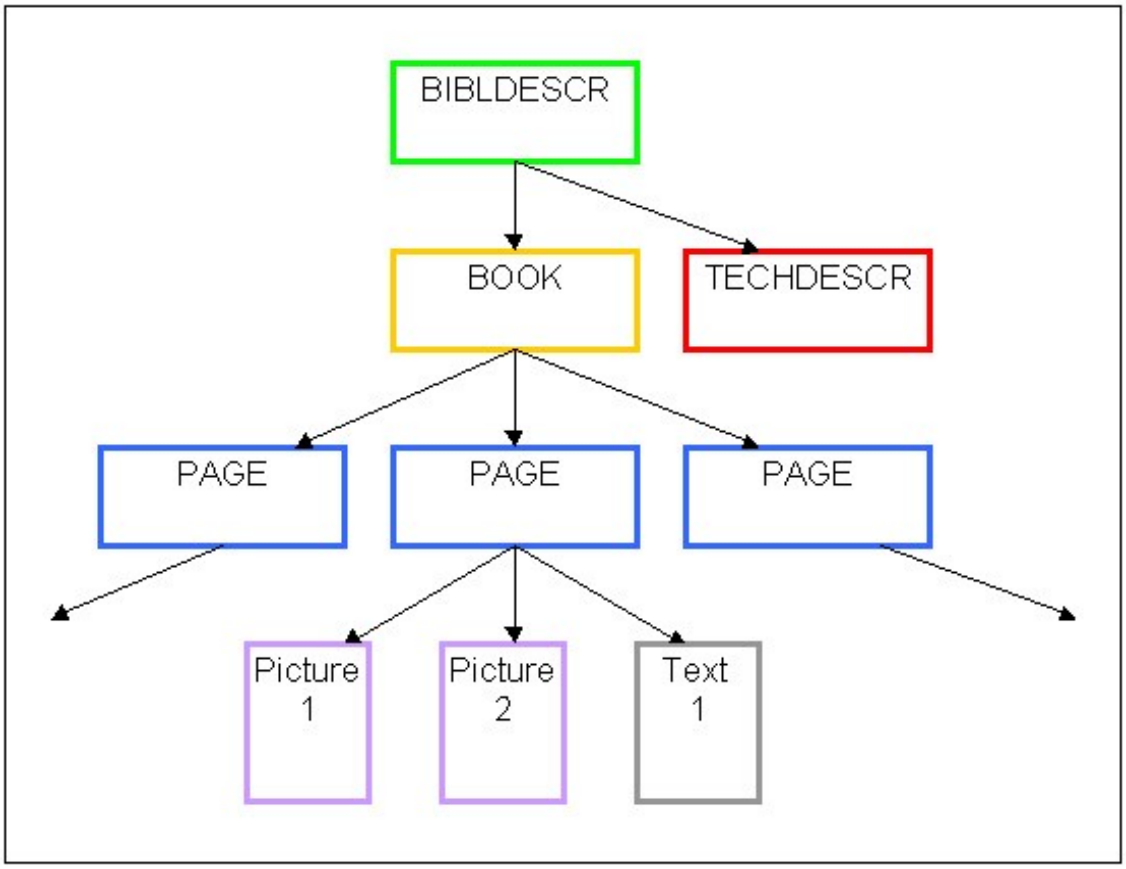
hybridní kamera a v roce 1999 byly vytvořeny metadatové standardy DOBM pro popis monografií a periodik, které vycházely z DOBM pro rukopisy. Vznikla vlastně specifikace typu dokumentu periodika a monografie pro obecný standard DOBM. DOBM se používalo do roku 2002, o rok později jej nahradila nová DTD postavená na XML. V roce 2000 bylo v rámci *Krameria* naskenováno již zhruba 100.000 stránek (v současnosti (2012) je to zhruba 9 milionů stran periodik a monografií). Probíhala souběžně digitalizace fyzických předloh, převážně periodik, a digitalizace mikrofilmů.
Metadatový popis periodik i monografií v DOBM byl v podstatě pouze bibliografický, doplněný o tři elementy, které lze považovat za technická metadata. Tyto tři elementy měly spíše informační charakter o technickém řešení, na kterém byly dokumenty digitalizovány, než že by měly za cíl dlouhodobou ochranu digitálních dat. Pokud by takový cíl tyto tři elementy měly, byly by zcela jistě doplněny dalšími podstatnými údaji, jako je např. datum a čas vzniku digitální kopie, kdo předlohu digitalizoval, velikost souboru apod. Tyto chybějící údaje lze považovat za administrativní metadata a jak se ukázalo při testování převodu DTD do interních formátů LTP systému v roce 2010, tento typ údajů v současnosti při dalším zpracování metadat velmi citelně chybí, zvláště v případě, že si nikdo okolnosti vzniku digitálních souborů již přesně nepamatuje.

Povinnou součástí DOBM záznamu monografií i periodik byl soubor ENTER.SGM. Obsahoval seznam elementů pro každý metadatový soubor (viz níže). Zároveň soubor ENTER.SGM (a v něm DOBMENT.DTD) označoval přímo kořenový soubor digitálního dokumentu [KNOLL, 2000, s. 2]. Komplexní digitální objekt reprezentující fyzickou monografii sestával z různých souborů, z vlastních dat a souvisejících metadat – viz Obrázek 16. Metadata pro monografie, podobně jako u rukopisů, byla rozdělena do následujících kategorií, každou z nich tvořil jeden DOBM záznam [KNOLL a VOMLEL, 1999a]:

- **BIBLDESCR** – bibliografický popis – obsahoval obdobné údaje jako katalogizační záznam, dále odkazy na část popisující technická metadata a na soubory tvořící monografii;
- **TECHDESCR** – technický popis – obsahoval údaje o procesu digitalizace, pouze jeden element <capturingData>, který mohl obsahovat nestrukturovaný text[192] uvádějící popis skeneru, na kterém proběhlo skenování, originální rozlišení nebo informace o formátech;
- **BOOK** – kniha – tento soubor reprezentoval fyzickou strukturu monografie (stránky, obálky apod.) včetně jejího vlastního obsahu (zdigitalizovaných souborů stránek); z tohoto souboru vedly odkazy na jednotlivé stránky v souboru PAGE;
- **PAGE** – stránka – tento soubor představoval jednu stránku knihy; soubor PAGE obsahoval náhled obrazu stránky i její popis; jeho součástí byly také odkazy na vlastní digitální soubory stránek – různé kvality obrazu; odkazy na textový soubor s OCR apod.

Metadata pro popis monografií ve standardu DOBM bohužel nedokázala vyjádřit logické vnitřní části, jako např. kapitoly, přílohy apod. Tato možnost byla přidána až v roce 2003 s přechodem na XML a vznikem nového standardu nazývaného DTD monografie – viz kapitola 5.2.2.2.

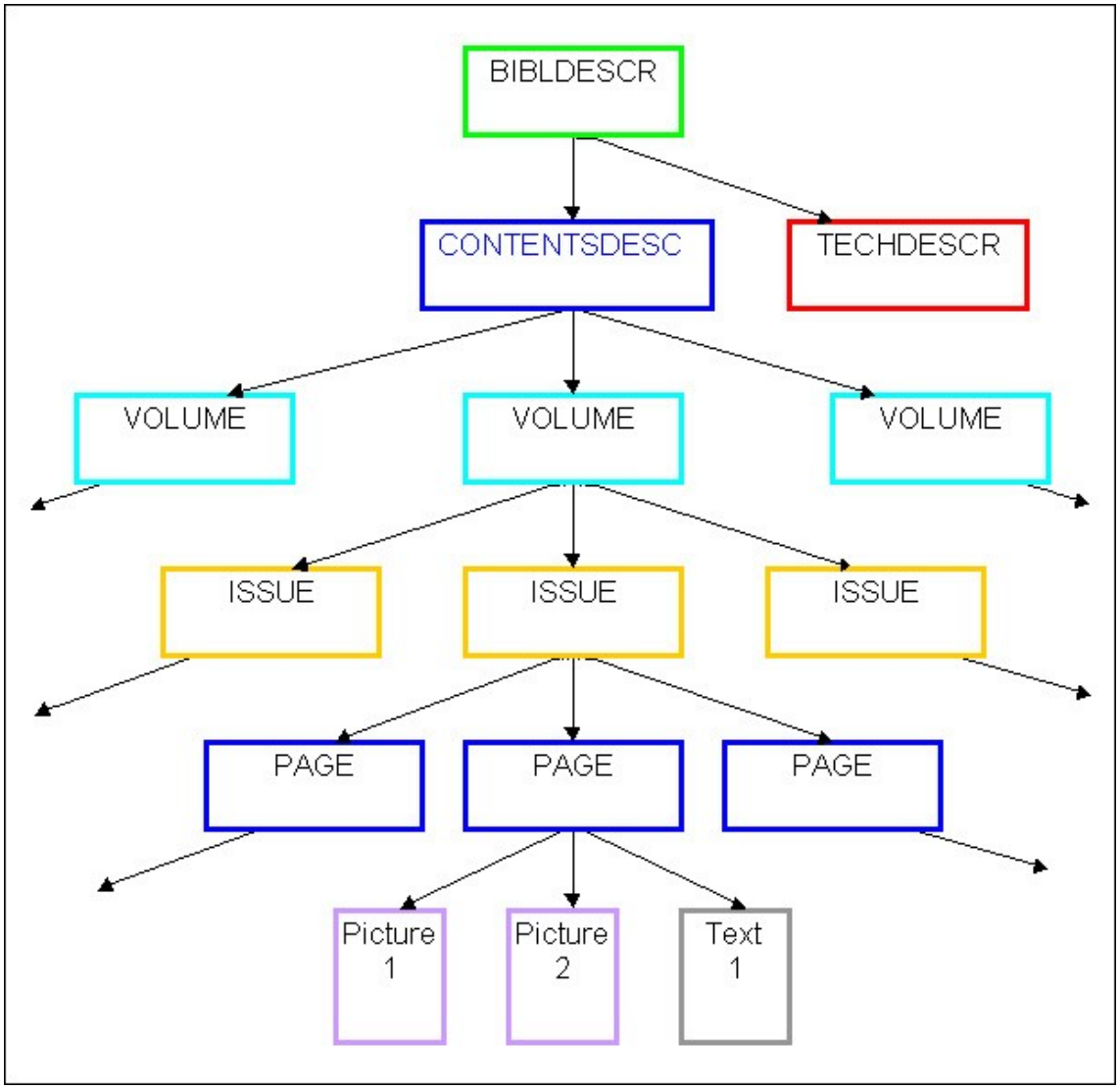
[192] Problémovost vyplývající z toho, že technická metadata nebyla strukturovaná, se přenesla i do budoucího DTD, kde sice byly pro technický popis použity tři elementy, ale i ty se plnily nestrukturovaným textem podle proměnných pravidel, často navíc nesprávně tak, že se hodnoty vyplňovaly předem pro určitý dokument, i když nebyly pravdivé (v případě, že se např. skener v půli skenování určitého titulu z důvodu poruchy změnil apod.).



Obrázek 16 – Struktura metadatového popisu DOBM SGML pro monografie [KNOLL a VOMLEL, 1999a].

DOBM metadata pro periodika byla rozdělena do následujících kategorií [KNOLL a VOMLEL, 1999c], kde stejně jako u monografií každou z nich tvořil samostatný DOBM soubor, jak ukazuje také Obrázek 17:

- **BIBLDESCR** – bibliografický popis – obsahoval katalogizační údaje o periodiku a odkazy na část technických metadat TECHDESCR stejně jako na soubory v části CONTENTSDESCR;
- **TECHDESCR** – technický popis – údaje o procesu digitalizace v rámci jednoho elementu <capturingData>, obsahoval stejně jako v případě monografií nestrukturované údaje;
- **CONTENTSDESCR** – popis struktury obsahu – tento soubor popisuje periodikum jako celek, obsahuje skupinu ročníků a linky na soubory ročníků;
- **VOLUME** – ročník – tento soubor reprezentoval jeden ročník, resp. sadu souborů, které jej tvořily (obsahoval také linky na soubory jednotlivých čísel nebo stránek);
- **ISSUE** – číslo – tento soubor reprezentoval jedno číslo, identifikoval vydání a také obsahoval seznam a popis článků a ilustrací i linky na jednotlivé stránky;
- **PAGE** – stránka – tento soubor reprezentoval jednu stránku periodika a mohl obsahovat linky na různé úrovně kvality obrazových souborů reprezentujících stránku včetně jejího popisu a OCR souboru v textové podobě; neobsahoval ovšem popis jednotlivých článků.



Obrázek 17 – Struktura metadatového popisu DOBM SGML pro periodika [KNOLL a VOMLEL, 1999a].

V metadatovém standardu DOBM pro periodika nebylo možné popsat článek, popis byl pouze na úroveň stránky. Tato možnost byla přidána v novém standardu nazvaném DTD periodika a ani poté nebyla z důvodů časové náročnosti využívána.[193]

Metadatový popis monografií i periodik se mohl rozvíjet i nad úroveň bibliografického popisu. Existovala možnost vytváření kolekcí (COLLECTION). Tento přístup byl použit při popisu sbírky arabských rukopisů, která je dostupná v digitální knihovně Manuscriptorium [KNOLL, 2000, s. 2].

5.2.2.2 DTD monografie, DTD periodika a jejich vývoj

Pro popis monografií a periodik v programu digitalizace *Kramerius* se zpočátku využívaly metadatové standardy vycházející z DOBM SGML periodika a monografie – viz kapitola 5.2.2.1. Nová specifikace standardů (DTD monografie a DTD periodika) vznikala v roce 2003 již na bázi XML a na DOBM SGML logicky navazovala. Obě DTD jsou z hlediska struktury DOBM velmi podobná. Autorem návrhu původního DOBM i finálních DTD byl Adolf Knoll. První DTD pro periodika dokončil na jaře 2003 (březen 2003 verze 1.00). DTD monografie dokončil na podzim

[193] Se zpracováním článků se počítá až v projektu NDK od roku 2012/2013.

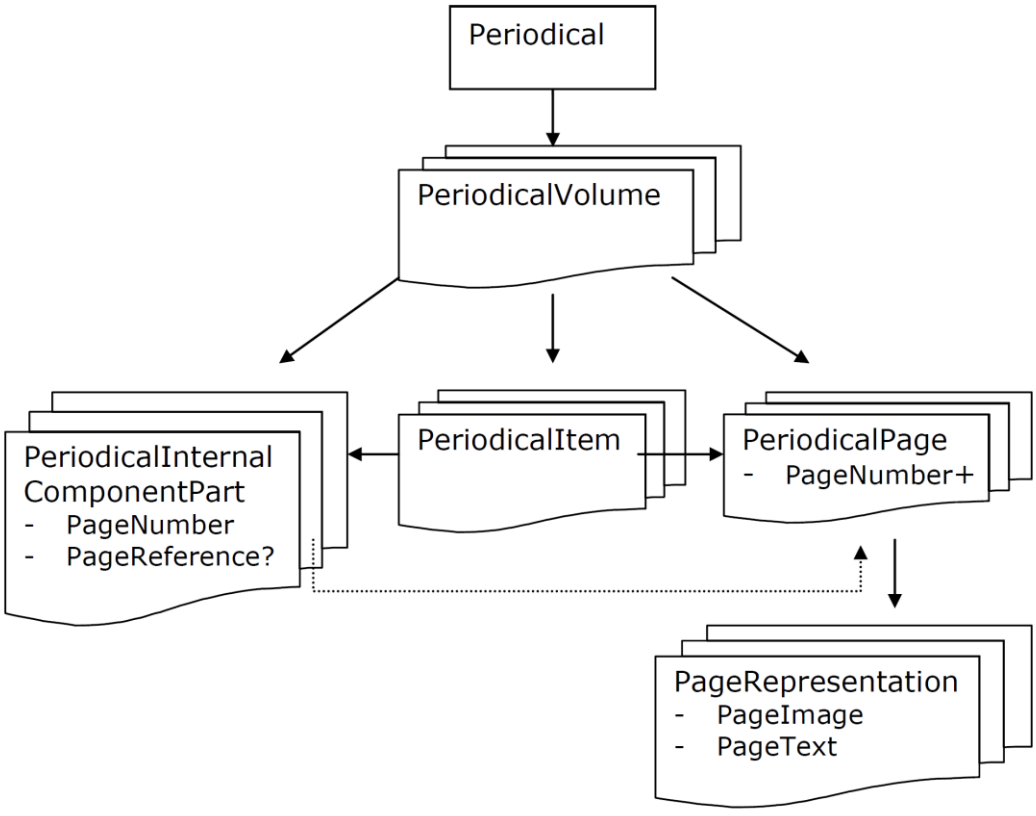
2003. Specifikace obou DTD vznikala přímo pro aplikaci zpřístupnění Kramerius a naopak, Kramerius systém byl od počátku uzpůsoben na práci se zmíněnými XML DTD standardy (což se později ukázalo jeho nevýhodou).
Kramerius ovšem nebyl první možností jak si prohlížet zdigitalizované novodobé dokumenty na Internetu. Periodika byla na webu NK ČR dostupná již dříve, a to přímo z digitálního archivu uloženém na páskovém robotu s *on-the-fly* konverzí do DjVu [KNOLL, 2011a]. Aplikace, která od roku 2001 umožňovala mj. přístup ke zdigitalizovaným rukopisům i novodobým dokumentům se nazývala AIP SAFE a pracovala s DOBM SGML pro periodika a rukopisy (více viz také kapitola 6.2). Koncept zpřístupňování historických dokumentů i novodobých dokumentů v jedné aplikaci nebyl ovšem dlouhodobě schůdným. Ukázalo se, že díky odchylkám při vyplňování konkrétních polí a odlišné specifikaci metadat u historických dokumentů a novodobých dokumentů, nebylo možné do aplikace AIP SAFE valnou většinu historických dokumentů vůbec naimportovat [KNOLL, POLIŠENSKÝ a UHLÍŘ, 2004, s. 32]. Díky tomu, že dokumenty i jejich potencionální uživatelé jsou natolik odlišné, byly v roce 2003 vytvořeny dvě různé aplikace, pro historické dokumenty (Memoria/Manuscriptorium) a pro dokumenty novodobé (Kramerius). Tyto aplikace pracovaly a dodnes pracují s XML metadaty. Standard DOBM SGML byl opuštěn a v aplikaci Kramerius nahrazen DTD pro periodika a pro monografie.[194] Metadata pro historické dokumenty v podobě XML šla jinou cestou – viz kapitola 5.2.1.2. Aplikace AIP SAFE nově vzniklá DTD na základě XML nepodporovala a data do ní nebyla dále ukládána. Migrace již hotových záznamů ve standardu DOBM SGML do obou DTD proběhla během jednoho dne. Celou migraci zajišťoval externí nástroj vytvořený pro NK ČR[195]. Ovšem ukázalo se, že mapování nebylo zcela přesné a následné ruční opravy takto vytvořených XML záznamů trvaly velmi dlouho.

Návrh obou DTD vycházel z již používaného standardu DOBM SGML pro periodika a monografie. V DTD byla více propracována problematika vnitřního členění monografií a periodik, včetně typů logických vnitřních částí a byly také doplněny bohatší možnosti vnitřní struktury (analytický popis na úroveň článků). Zároveň bylo upuštěno od systému několika úrovní kvalit obrazových reprezentací stránky, které vycházely z praxe běžné v digitalizaci a digitální knihovně Manuscriptorium (Gallery Quality Picture; Preview Quality Picture; Normal Quality Picture; Internet Quality Picture; Excellent Quality Picture; Detail Quality; Watermark). Z těchto několika kopií zůstala pouze archivní a uživatelská kopie.
Část DTD pro popisná metadata vycházela ze standardu MASTER, který se od roku 2003 používal pro vyjádření popisných metadat v projektu *Manuscriptorium* [RÁKOCY, 2008, s. 13]. Původně se počítalo s tím, že bude při specifikaci DTD pro periodika využito výsledků projektu DIEPER, který probíhal mezi lety 1998-2000 jako evropský projekt ve čtvrtém rámcovém programu (FP4)[196]. Na

[194] Aplikace pro zpřístupnění Kramerius měla původně poskytovat přístup také k digitalizovaným muzejním objektům a zvukovým dokumentům. K tomu ovšem nikdy nedošlo, i přesto, že příslušná DTD byla připravena – viz [KNOLL, 2003, s. 235]. DOBM formát pro zvukové dokumenty a muzejní objekty byl převeden na XML a do DTD během roku 2004 [KNOLL, 2011a].
[195] Tento nástroj zároveň zajišťoval migraci původních JPEG používaných v AIP SAFE do formátu DjVu, který se jevil jako vhodnější pro rychlejší zpřístupnění digitalizovaných periodik.
[196] Cílem projektu DIEPER bylo primárně vybudování mezinárodní databáze digitalizovaných periodik s permanentními URL a permanentním přístupem k dokumentům. Hlavním hlediskem byla ochrana papírových předloh a koordinace projektů digitalizace. Od tohoto se odvíjí snaha vytvořit metadatový formát pro digitalizovaná periodika.

základě analýz[197] a porovnání bylo konstatováno, že DIEPER odpovídá potřebám NK ČR v oblasti popisu periodik – viz [KNOLL, 2002, s. 38], i když přístup projektu DIEPER k metadatům se od Krameria poněkud lišil. Např. tím, že DIEPER tíhnul k popisu vědeckých článků. Podobné praxi v Krameriovi bylo naopak členění vnitřních částí na ročníky, čísla a články. Podstatou jednoznačné identifikace těchto částí v návrhu DIEPER byl identifikátor SICI [KNOLL, 2002, s. 37]. SICI identifikátor byl poté převzat i do specifikace DTD pro Krameria, ovšem nebyl nikdy plněn. Bohužel ale nedošlo během projektu DIEPER k vytvoření finální verze DTD, pouze k slovnímu popisu polí a struktury a proto nebylo možné standard přebrat. I tak se ovšem některé prvky v novém DTD pro periodika objevily, např. struktura popisu, popisy článků apod. Finální DTD pro periodika tak vzniklo na základě původního popisu ve standardu DOBM, na základě analýz návrhů nového standardu v projektu DIEPER a na základě zkušeností a potřeb.

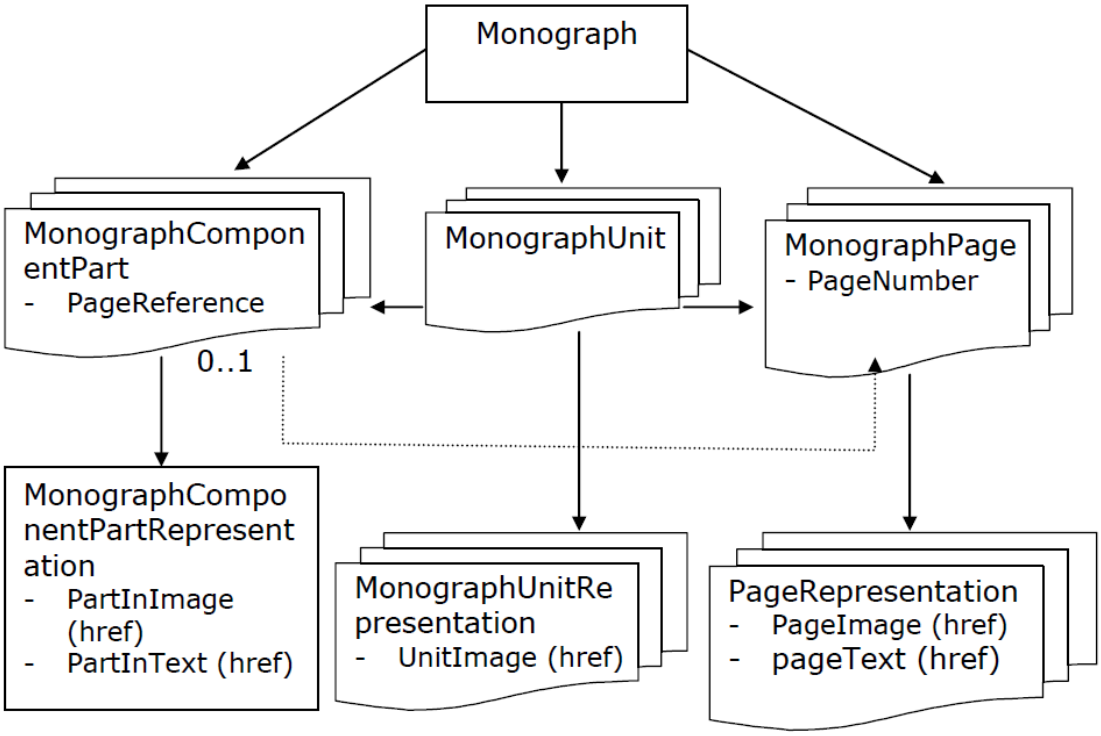
Vzhledem k tomu, že v době, kdy vznikaly standardy DTD periodika a monografie, neexistoval specializovaný metadatový standard pro popis struktury dokumentu, byla tato DTD navržena tak, aby obsahovala popisná metadata i popis struktury. Záznamy tak vyjadřovaly strukturu popisovaného titulu periodika – viz Obrázek 18, který ukazuje titul periodika skládající se z ročníků (PeriodicalVolume). Ročník může obsahovat čísla (PeriodicalItem) a také na čísle nezávisle vnitřní části (PeriodicalInternalComponentPart), což jsou např. články, předmluva aj., nebo přímo stránky (PeriodicalPage). Poslední dvě jmenované entity může samozřejmě obsahovat i číslo periodika. Popisná metadata jsou na úrovni titulu, ročníku, čísla, vnitřní části i stránky. Jednotlivé údaje lze v různých úrovních struktury dědit. Podrobnost metadat se na různých úrovních liší. Z popisu stránky je odkaz na vlastní data (PageRepresentation), což je naskenovaný obraz a případně OCR.



Obrázek 18 – Zjednodušený pohled na strukturu DTD periodika [QBIZM TECHNOLOGIES, 2007, s. 39].

[197] http://digit.nkp.cz/EnglishArticles/DOBM_DIEPER.xls

Podobně je to u monografií – viz Obrázek 19. Monografie obsahuje svazek (MonographUnit) anebo přímo vnitřní části (MonographComponentPart) a stránky (MonographPage), pokud se svazek vynechá (jednosvazková publikace). Svazek samotný také může obsahovat vnitřní části a stránky. V případě monografií může mít své vyjádření (naskenovanou stránku, OCR) i vnitřní část, např. pokud jde o tabulku, fotografii apod. Stejné je to i pro svazek (MonographUnit), kterým může být další tabulka, mapa, příloha aj. Metadatový popis je na úrovni titulu, svazku, vnitřní části i stránky. Stejně jako v případě periodik se určité údaje mohou dědit a jednotlivé úrovně se v podrobnosti popisu liší.



Obrázek 19 – Zjednodušený pohled na strukturu DTD monografie [QBIZM TECHNOLOGIES, 2007, s. 27].

Vytváření metadat v nové podobě DTD periodika bylo možné pomocí nástroje Sirius[198] od firmy Elsyst Engineering (Vyškov), který byl upraven tak, aby dokázal pracovat s XML a s konkrétními DTD definicemi typu dokumentů pro periodika i monografie. Sirius podstatně urychlil tvorbu metadat na rozdíl od tvorby metadat ve standardu DOBM SGML v aplikaci AIP SAFE – více viz [KNOLL, POLIŠENSKÝ a UHLÍŘ, 2004, s. 22]. Sirius je dodnes nejrozšířenějším nástrojem na tvorbu metadat v DTD periodika a monografie v ČR.

DTD monografie i periodika obsahovala technická metadata. Ovšem v podobně minimalistické verze jako byla metadata vyjádřená v původním DOBM SGML. Cílem technických metadat bylo udržet informaci o skeneru a nastavení skenování. Administrativní metadata ve smyslu kdo, kdy a kde obrazy a metadata vytvořil, chyběla v obou DTD specifikacích zcela. Technický popis již od verze DTD 1.0 byl obsažen v následujících elementech:

[198] http://www.ee.cz/sirius/index.htm

<TechnicalDescription>
<ScanningDevice> – povinný element
<ScanningParameters> – nepovinný element
<OtherImagingInformation> – nepovinný element

Během let se tyto elementy plnily různě, od volného textu až po strukturovaný volný text dle pravidel popisu (monografií i periodik). U monografií zpracovaných dle pravidel popisu verze 1.5 z 1. 10. 2009 mohou tyto elementy nabývat následujících hodnot [NÁRODNÍ KNIHOVNA ČR, 2009, s. 7]:

- <ScanningDevice>
  a) druh digitalizace – přímá, z mikrofilmu, z mikrofiše;
  b) typ skeneru (mikrofilmový, knižní, plochý, bubnový aj.);
  c) model skeneru;
  d) výrobce skeneru;
  e) použitý SW;
  f) verze SW;
  g) výrobce SW
- element <ScanningParameters>
  a) rozlišení: nejdřív číslo, mezera a DPI; pokud je identická osa X i Y, uvádí se pouze jedno číslo
  b) barevná hloubka uvedená ve tvaru počtu bitů
  c) barevná škála (BW – GS – RGB)
- element <OtherImagingInformation>
  a) uvede se verze pravidel popisu, podle které se dokument zpracovával

Stejně technická metadata vypadají u periodik, dle posledních pravidel popisu z roku 2008 [NÁRODNÍ KNIHOVNA ČR, 2008, s. 3]. Technický popis, tj. element <TechnicalDescription> bylo možné přidat ke každé reprezentaci všech úrovní monografie (vrchní úroveň titulu, úroveň svazku, vnitřní části, stránky). Podobné to je u periodik, kde technický popis lze přidat k titulu, ročníku, číslu, vnitřním částem a stránce. Nejčastěji byla technická metadata přidávána přímo u popisu reprezentací jednotlivých stránek. Někdy také pro celou popisovanou entitu (ročník, číslo; monografie), pokud byly všechny obrazy entity skenovány na jednom skeneru se stejným nastavením.

Pomocí DTD periodika bylo a je možné popsat např. celý titul časopisu, včetně všech ročníků a čísel. Stejně tak je možné popsat např. pouze jeden ročník, nebo více ročníků. Teoreticky by šlo popsat i jediné číslo. V realitě je ovšem popis monografií a periodik vázán na fyzický svazek, který se digitalizuje. Tento přístup může být v pořádku v případě rukopisů a starých tisků, kde se často vyskytovaly v jednotlivých svazcích tzv. přívazky, a bylo potřeba popsat celý svazek. Problém tento přístup nepůsobil ani v případě jednosvazkových novodobých monografií, pro jednu knihu jako intelektuální entitu vzniká jeden metadatový popis. Velkým problémem do budoucnosti, který nikdo nepředvídal a ukázal se při testování LTP systémů v roce 2010, je skutečnost, že se tímto způsobem digitalizovaly/popisovaly i fyzické svazky periodik. V knihovnické praxi je běžné, že jsou periodika svazována do jednoho svazku např. po jednom ročníku, pokud je čísel do roka příliš mnoho, váží se např. po měsících, půlrocích nebo jen dle potřeby. Pokud takto vzniklý svazek

projde bez dalšího zamyšlení digitalizací, na konci vznikne jeden metadatový záznam např. pro šest čísel periodika (leden–červen) a druhý záznam pro další svazek pro čísla červenec–prosinec. Vzniknou tedy záznamy nekonzistentní mezi sebou, někdy s popisem ročníku, někdy s popisem půlky ročníku, někdy také s popisem čísel vyšlých během více let. To v určitých případech komplikuje další práci s metadatovým záznamem. Často vznikal metadatový záznam, který byl neúměrně dlouhý (desetitisíce řádků). To nastávalo zejména v případech, kdy šlo o deník a metadatový popis fyzického svazku jednoho ročníku obsahoval popis několika set čísel a linky na ještě větší počet souborů (Reichenberger Zeitung, Národní listy aj.).[199] Nekonzistence intelektuálních entit působí potíže při jakémkoliv importu metadat do jiného systému, než je např. Kramerius. Při takovém importu musí vždy dojít k rozpojení/spojení metadatových záznamů, nebo k jejich demontáži na jednotlivé intelektuální entity (čísla např.).

Intelektuální entita by neměla být nahodilá, ale daná (např. číslo, ročník). Ve velkých projektech digitalizace ve světě je dnes základní popisovanou intelektuální entitou číslo periodika, ke kterému vzniká metadatový záznam obsahující i údaje o vyšších úrovních (ročník, titul aj.). Základní intelektuální entitou pro popis monografií je svazek monografie. Takto je nastavena digitalizace v projektu NDK a bude tak velmi pravděpodobně nastaven i program VISK7 od roku 2013.
V případě digitalizace a popisu na čísla nebo svazky, vyjadřují strukturální metadata záznamu čísla pouze pořadí stránek ve výtisku čísla (nebo svazku monografie) a jejich logické části. Strukturální metadata pro celý titul periodika nevznikají. Spoléhá se na to, že moderní systémy zpřístupnění digitálních dat (digitální knihovny) jsou schopné agregovat uživateli výsledky dle jeho dotazu tak, že uživatel uvidí všechna čísla, roztříděné na ročníky např. ve stromové struktuře chronologicky, pokud bude hledat např. podle názvu titulu periodika. K tomu je samozřejmě nutné dodržovat určité podmínky metadatového popisu, uvádět tituly a názvy podle autorit, uvádět důsledně všechny dostupné identifikátory (čČNB, ISSN, ISBN apod.) vrchních entit (titul) u popisu čísel nebo svazků monografií. Nemusí tak docházet, jak se dělo např. v Krameriovi, k umělému slučování několika svazků do jednoho titulu, které často nemá oporu ani v katalogizačním záznamu.

5.2.3 Metadata z pohledu dlouhodobé ochrany digitálních dat v NK ČR v období 1996-2003

První období digitalizace bylo v NK ČR, ale i jinde ve světě, ve znamení popisných metadat a výzkumu v oblasti digitalizace. V České republice ani v NK ČR se proto aktivně logická dlouhodobá ochrana digitálních dat dlouho neřešila. Řešitelům projektů *Kramerius* i *Manuscriptorium* ovšem bylo jasné, že v budoucnu bude nutné stávající obrazové i jiné soubory převádět na nové formáty, aby byly dále použitelné. Budoucnost se ovšem zdála příliš daleká a žádné konkrétní aktivity směřující k logické dlouhodobé ochraně dat neprobíhaly. Určitý důraz byl kladen na používání standardů dat i metadat, ovšem spíše jen na striktní používání konkrétních specifikací (DOBM SGML, MASTER+, DTD periodika a monografie), které zajistí konzistenci dat i metadat a také definici způsobu popisu dokumentů. Za důležitou v tomto ohledu byla považována i striktní formalizace zápisu konkrétních polí metadat, což se např. v digitalizaci novodobých dokumentů ne vždy dařilo (nekonzistence údajů o odpovědnosti). Pouhé používání určitých standardů nemusí

[199] Pro aplikaci zpřístupnění, v tomto případě Kramerius do verze 3, docházelo ke spojování všech metadatových záznamů jednoho titulu periodika do jednoho XML záznamu, který pak mohl mít i více než sto tisíc řádků a velikost přes 5 MB.

stačit, pokud se nejedná o mezinárodně přijímaný a používaný standard, což byl problém hlavně u obou výše zmíněných DTD. Ta byla sice postavena na standardu XML a definována standardním DTD souborem, ale jako celek, co do obsahu, byla proprietární. Je nutné uznat, že do roku 2003 neexistovalo mnoho metadatových standardů, které by bylo možno přebrat.

Nebylo přihlíženo cíleně k ochraně obrazových souborů. V DOBM SGML byla informace o technických vlastnostech digitálních obrazů naprosto minimální. Rozsáhlejší technická a do jisté míry i administrativní metadata vznikala v digitalizaci historických dokumentů až se standardem MASTER+, ovšem za účelem, kterým bylo v případě projektu *Manuscriptorium* maximálně věrné zobrazení uživateli. Chyběly tak např. kontrolní součty, velikost jednotlivých souborů aj. Pozitivní je použití specifikace NISO pro obrazové digitální dokumenty, která předznamenala standard MIX. Bohužel i tak v MASTER+ bylo plněno pouze několik polí technických metadat a to nejčastěji na vrchní úrovni titulu a ne jednotlivých digitálních objektů. V prostředí digitalizace novodobých dokumentů v rámci DOBM i pozdějšího DTD se také vytvářela určitá technická metadata. Obsahovala velmi omezený počet údajů, navíc v nestrukturované formě. Údaje popisovaly způsob digitalizace (přímá nebo z mikrofilmu); typ skeneru (výrobce, typ a použitý SW) a také o nastavení skeneru (rozlišení, barevná hloubka a barevná škála). To je z hlediska potřeb dlouhodobé ochrany nedostačující. Chybějí základní administrativní údaje, jako je datum skenování, datum tvorby metadat, odpovědnost za digitalizaci i za tvorbu metadat. Často chybějí bližší technická metadata pro jednotlivé obrazy a další digitální objekty, jsou uvedena obecně pro všechny na vrchní úrovni. Tento přístup trval až do roku 2008, kdy byly vytvořeny specifikace administrativních metadat a ta se začala v digitalizaci projektu *Kramerius* plnit – viz kapitola 5.3.2. Neexistence technických a administrativních metadat v období do roku 2003 a potažmo do roku 2008 a u určitých hodnot až dodnes, bude působit potíže při plánovaném přesunu archivních dat do plánovaného LTP systému v roce 2012/2013. Již při testování v roce 2010 jsme tyto údaje postrádali.[200]

Celkovou situaci v letech 1996-2003 lze přičíst stavu poznání v oblasti dlouhodobé ochrany, které bylo velmi omezené.

5.3 Národní knihovna ČR – období druhé (2004-2011)

Druhé období vývoje metadat v NK ČR lze charakterizovat zvládnutým procesem digitalizace a nastavenými standardy metadat v prvních letech. Po roce 2005 se ukazují nově dva aspekty, které jsou pro následující léta typické. Prvním z nich je skutečnost, že ve světě a později i v digitalizačních programech NK ČR se začínají využívat obecně rozšířené mezinárodní standardy metadat (MASTER, METS, PREMIS, MIX, MODS aj.). Tato změna nepostihla ani tak projekt *Memoria* a jeho digitální knihovnu Manuscriptorium, která standardní formáty používá od počátku, jako spíše projekt *Kramerius*. Velká část z nových standardů jsou standardy metadat pro ochranu – viz kapitola 4.8.2.

Druhým aspektem je zvětšující se zájem o problematiku dlouhodobé ochrany digitálních dat. Ta byla i jednou ze součástí VaV *Vytvoření virtuálního badatelského prostředí pro zpřístupnění a ochranu digitálních dokumentů* (2004-2010). V tomto období dochází v NK ČR k prvním úvahám o nástrahách logické dlouhodobé ochrany dat a také k pokusům o vytváření technických a

[200] V LTP systému tak bude nutné např. u data vzniku plnit datum vstupu do LTP systému nebo pouze nějaký odhad. Technická metadata na druhou stranu budou vytvořena v modulu Ingest. Bylo by ovšem dobré je validovat oproti hodnotám již v hotových metadatech, které s výjimkou kontrolního součtu, v DTD nejsou. Kontrolní součet chybí u historických dokumentů.

administrativních metadat s cílem zajištění předpokladů logické dlouhodobé ochrany digitálních dat. V roce 2008 byla v aplikaci Kramerius v základní podobě implementována metadata ve standardech PREMIS, MIX a METS, což byl významný krok kupředu. Historické fondy v této aktivitě zůstaly pozadu a technická a administrativní metadata pro ochranu dat PREMIS, MIX do svých procesů nepřevzaly i přesto, že standard metadat se u nich také měnil (ENRICH TEI P5). V roce 2006 v NK ČR vznikl referát a později v roce 2008 celý odbor digitální ochrany, který ve spolupráci s pracovištěm digitalizace novodobých dokumentů začal navrhovat implementace prvních ochranných metadat s cílem dlouhodobé ochrany digitálních dat.[201]

Celé období je charakteristické také zapojením do evropských projektů, které vyžadovaly zasílání a sdílení metadat, jako byl mj. projekt *The European Library* a jeho předchůdci i nástupci (EDLproject, TEL-ME-MOR), což NK ČR nutilo se zamyslet na interoperabilitou produkovaných metadat a osvojení světových standardů. Další projekty znamenaly pro NK ČR významný posun v oblasti metadat i dlouhodobé ochrany dat. NK ČR se účastnila, z hlediska knihovny samotné, zlomového projektu *DigitalPreservationEurope* ze šestého rámcového programu (FP6). Podobně zlomový, pro historické dokumenty, využití nových standardů a agregaci metadat, byl projekt ENRICH (2007-2009), také ze šestého rámcového programu (FP6), který NK ČR dokonce koordinovala – další informace o projektu viz [UHLÍŘ, 2010].

Konec druhého období je příznačný plánem na implementaci nových metadatových standardů pro digitalizaci novodobých dokumentů, a to v rámci projektu *Národní digitální knihovna*.

5.3.1 Metadatový popis historických dokumentů

V prvních letech neprobíhaly v oblasti metadat historických dokumentů žádné podstatné změny. V roce 2003 nasazený standard MASTER+ zcela vyhovoval pro uložení i prezentaci digitálních dokumentů v Manuscriptoriu. Problematickým se ukazoval způsob tvorby katalogizačních záznamů. Ty vznikaly v procesu digitalizace, resp. před ním, a to přímo ve standardu MASTER. Záznam se nedostal do katalogu NK ČR, byl pouze v Manuscriptoriu, které tak sloužilo jako katalog historických dokumentů. To ovšem nechtěly akceptovat ostatní knihovny, které tento proces dlouhodobě kritizovaly. Chtěly katalogizovat v MARC21, bylo to jednodušší a záznam bylo možné dát do běžného knihovního katalogu. Problém převodu mezi standardy UNIMARC/MARC21 a MASTER se řešil od roku 2003 a byl i tématem workshopu s názvem *Manuscriptorium: příprava dat a využívání informace*, který proběhl v rámci konference *Inforum 2005* – viz [IKAROS, REDAKCE, 2005]. Cílem byl import dat z katalogu NK ČR do katalogu Manuscriptoria i přímo do digitální knihovny. To mělo ulehčit tvorbu metadat v digitalizaci[202] a zajistit zapojení více institucí do digitální knihovny Manuscriptorium a jejího sdíleného katalogu historických dokumentů. Tento postup částečně řešil problém i pro knihovny, které historické dokumenty katalogizovaly v MARC21. Aplikace na tento převod byla vyvíjena v letech 2003-2004 [KNOLL, 2003, s. 238].

[201] Technická metadata vznikala např. v digitalizaci historických dokumentů již od roku 2003, ale nikoliv z důvodů dlouhodobé ochrany digitálních dat.

[202] Tento přístup je jedním z doporučovaných, zaručuje totiž mj. i to, že záznamy v katalogu a digitální knihovně budou totožné. Import bibliografických metadat z katalogu do metadat digitálního dokumentu umožňuje dnes většina nástrojů na workflow digitalizace nebo na tvorbu/kompletaci digitálních objektů – tj. dat a metadat. Jmenujme Sirius pro Krameria, M-tool pro Manuscriptorium a další komerčně dostupné SW, např. DocWorks od firmy CCS, Německo.

5.3.1.1 Standard METS v aplikaci Manuscriptorium

Rokem změn byl rok 2005. Uvažovalo se o zavedení stále více rozšířenějšího standardu METS, který podobně jako MASTER+ je kontejnerem a balí různé typy metadat. Standard METS, u kterého proběhla v roce 2005 analýza možného použití, byl uznán další možností, vedle MASTER+, jak vyjádřit komplexní digitální dokument určený pro archivaci [KNOLL, POLIŠENSKÝ a UHLÍŘ, 2005, s. 4]. V roce 2006 probíhaly práce na specifikaci a způsobu využití XML schématu Manuscriptoria v METS záznamu. Došlo k namapování standardu MASTER+ (msnkaip.xsd) do METS a vzniklo tak nové XML schéma pro komplexní digitální dokument. Jako vnitřní formát databáze Manuscriptoria byl METS ovšem zaveden až v roce 2007 v druhé verzi systému Manuscriptorium. Tento přechod byl jedním z úkolů projektu ENRICH. METS ovšem nikdy nebyl používán jako archivní metadatový standard, pro archivní kopie byl nadále používán standard MASTER+.

Jedním z důvodů přechodu na standard METS pro dokumenty v digitální knihovně Manuscriptorium byl nárůst počtu dokumentů v systému a nutnost údržby a správy metadat, ke kterým začaly, vedle popisných, přibývat ještě plné texty a jejich metadatový zápis [KNOLL, POLIŠENSKÝ a UHLÍŘ, 2007, s. 11]. Tato skutečnost si žádala nový pohled na digitální dokument. Řešením se ukázala být náhrada MASTER+ za standardní, obecně rozšířenější a univerzálnější standard METS – viz kapitola 7.2.1. Jeho výhodou je, že elegantně do sebe dokáže zapouzdřit různé typy různých metadat, např. i dva záznamy popisných metadat v různých schématech, což MASTER+ neumožňoval. Vzhledem k plánovanému přebírání metadat z mnoha knihoven světa v rámci projektu ENRICH, byla implementace METS standardu logickým rozhodnutím. METS nabídl Manuscriptoriu standard vhodný na výměnu (jakýchkoliv) metadat.

Využití METS pro komplexní digitální dokument kopírovalo zaběhnutou praxi v projektu *Manuscriptorium*, tedy podoba vložených metadat vycházela ze standardů MASTER a MASTER+. Veškeré popisné údaje byly v části <dmdSec> ve standardu MASTER nebo v jiném, technická a administrativní metadata pak v části <amdSec> a jejích podčástech. Pro technická metadata byl pro obrazové soubory doporučen standard MIX [AIP Beroun, 2005, s. 38] také proto, že původní technická metadata používaná v prostředí Manuscriptoria vycházela ze stejného základu jako standard MIX a mapování bylo tedy jednoduché[203] – viz kapitola 7.2.4. Technická metadata audio a jiných dokumentů budou uložena ve stejné části, tj. v <techMD>, ale v jiných schématech. Údaje o fyzické předloze byly v části <sourceMD>, která obsahuje popisy stránek fyzické předlohy, které mohou být velmi strukturované i dlouhé. Využívána byla i část <rightsMD> pro zápis údajů o autorských právech např. pro obrazy, pro plné texty dokumentů. Strukturální mapa (může jich být i více) byla v části METS <structMap>. Veškeré soubory tvořící komplexní objekt byly pak popsány v části <fileSec> v jednotlivých <fileGrp> např. pro XML soubory plných textů, různé kvality obrazů, audio záznamy aj. Podrobný popis komplexního digitálního dokumentu viz [AIP Beroun, 2005, s. 14 a dále].

Vzhledem k cílům digitální knihovny Manuscriptorium a VaV *Vytvoření virtuálního badatelského prostředí pro zpřístupnění a ochranu digitálních dokumentů*, došlo k tomu, že důraz byl kladen na uživatelské rozhranní digitální knihovny Manuscriptorium a lze říci, že veškerý vývoj v oblasti

[203] Technická metadata byla stále považována primárně za způsob zachycení údajů o digitalizaci, méně pak o digitálním objektu samotném.

metadat byl podmíněn potřebami a trendy této digitální knihovny – viz např. [KNOLL, POLIŠENSKÝ a UHLÍŘ, 2008, s. 3]. Většina vývoje v oblasti metadat se tomu podřídila, byly vytvářeny a přejímány metadatové standardy vhodné pro sdílení metadat a jejich výměnu. Standard metadat, který vznikal při výrobě a byl používán i při archivaci se od přechodu na MASTER+ nezměnil. S přechodem na TEI P5 ENRICH jako výrobní standard i archivační standard se počítá až na rok 2012. I standard METS, který se velmi hodí pro uložení komplexních digitálních objektů, byl používán jen pro potřeby digitální knihovny Manuscriptorium a tedy pro uložení dat a metadat potřebných pro zpřístupnění.

5.3.1.2 TEI P5 ENRICH

Podstatné změny přinesl pro Manuscriptorium evropský projekt ENRICH. V jeho rámci bylo od roku 2007 Manuscriptorium platformou pro zpřístupnění a tedy sdílení metadat z různých digitálních knihoven Evropské unie i dalších států. Již na začátku projektu se začalo s vytvářením metadatového standardu postaveného na TEI P5, který byl později nazván ENRICH.DTD. ENRICH standard TEI P5 vycházel mj. také z práce, kterou udělalo konsorcium TEI. TEI P5 totiž obsahuje novou část věnovanou pouze popisu rukopisů. Sada elementů, někdy pod názvem TEI P5 MS, je založena na formátu MASTER a také na práci *TEI Medieval Manuscripts Description Work Group* (TEI-MMSS) [DRISCOLL, 2006]. Od počátku byl standard ENRICH plánován tak, aby byl zcela v souladu se specifikací TEI P5.

V rámci projektu ENRICH bylo několik pracovních skupin a úkolů (*workpackage* – WP), které se z různých hledisek zabývaly metadaty. WP3, který se zabýval standardizací metadat mj. měl za úkol vytvořit konverzi mezi platformami TEI P4 a TEI P5 pro popis rukopisů a rozšíření interního prostředí digitální knihovny Manuscriptorium o kontejnerový standard METS. První jmenovaný úkol představoval přechod ze standardu MASTER a MASTER+ ke specifikaci TEI P5 ENRICH. Instituce zapojené v projektu ENRICH se vrátily ke standardu MASTER (TEI P4) a na jeho základě navrhly nový standard na popis rukopisů na bázi TEI P5, který umožňuje vytvořit komplexní digitální dokument s popisnými metadaty i odkazy na vlastní data a vyjádřením struktury. Jak MASTER+, tak i TEI P5 umožňují vytvářet metadatový záznam obsahující různé typy metadat (např. popisná, technická i strukturální). TEI P5 ENRICH (ENRICH.DTD) přizpůsobil popis i potřebám prvotisků a starých tisků, dovoluje tak vytvářet pro tyto typy dokumentů popisná metadata přímo a dle [UHLÍŘ, 2010, s. 11] svými možnostmi překonává „*archaičnost a antikvovanost MARCových formátů*“. Podobně také [KNOLL, POLIŠENSKÝ a UHLÍŘ, 2004, s. 7].[204]

TEI P5 ENRICH je na bázi TEI P5, což znamená, že využívá jeho struktury, zásad i elementů s tím, že v několika částech přidává elementy nové a v různých částech zase elementy TEI P5 nepoužívá. „*Do schématu ENRICH jsou zahrnuty čtyři klíčové moduly TEI - header, core, tei, a textstructure*.[205] *Zároveň bylo přidáno pět speciálních modulů: msdescription, linking, namesdates, figures, a transr.*“ [BURNARD, CUMMINGS a RAHTZ, 2008, s. 5] Modul „*msdesription*“ obsahuje bibliografický popis předlohy, jde o nejdůležitější část u popisu historických dokumentů (element <msDesc>). Modul „*linking*“ popisuje linkování v textu, jeho segmentaci na různé části nebo také souběhy paralelních textů apod. Modul „*namedates*“ obsahuje popis jmen, osob, míst a dat a modul „*figures*“ obsahuje popis grafů, tabulek. Konečně poslední modul použitý ve specifikaci

[204] Bohužel, klasické knihovní katalogy stále fungují na standardu MARC21, takže přínos pro ně z tohoto pohledu je nepatrný.

[205] Tyto moduly by měly být součástí každého TEI záznamu, ať popisuje jakýkoliv typ dokumentu.

ENRICH „*transr*“ obsahuje transkripci textů. Byl také použit modul „*gaiji*“ popisující způsoby zápisu nestandardních znaků.

ENRICH.DTD a vlastně celý popis komplexního digitálního dokumentu v Manuscriptoriu má tři hlavní části:

- <teiHeader> – popisuje fyzickou předlohu digitalizovaného dokumentu, veškerá popisná metadata;
- <facsimile> – popisuje digitalizované obrazy a strukturu celého komplexního dokumentu;
- <text> – obsahuje libovolný text, využívá se pro přepis plného textu dokumentu.

První dvě části jsou potřebné pro digitální knihovnu Manuscriptorium, třetí část je volitelná. V první části popisující fyzickou předlohu digitálního dokumentu, jsou ze čtyř možných částí elementu <teiHeader> použity pouze dvě, a to element <fileDesc> obsahující bibliografický popis digitálního objektu, a <revisionDesc> zachycující všechny změny v záznamu. Element <fileDesc> obsahuje vnořené elementy <titleStmt>, <publicationStmt> a <sourceDesc>. Poslední z nich popisuje zdroj, ze kterého vznikl digitální dokument, tj. obsahuje bibliografický popis fyzické předlohy v elementu <msDesc>. Popis vychází z TEI P4 a tedy i ze standardu MASTER, se kterým sdílí většinu elementů z jeho části <msDescription>.
Další část <facsimile> s metadaty popisujícími digitální dokument je logicky pouze u komplexních dokumentů digitalizovaných. Obsahuje element <surface> pro každou stránku s vnořeným elementem <graphic> pro všechny digitální obrazy náležející k té konkrétní stránce. Specifikace technických metadat pochází přímo z TEI P5 standardu.

Digitální knihovna Manuscriptorium[206] přešla na standard TEI P5 ENRICH (ENRICH.DTD) v roce 2009, tedy v posledním roce řešení projektu ENRICH. Do té doby využívala aplikace Manuscriptorium standard MASTER+. TEI P5 ENRICH funguje dodnes jako její interní formát i formát pro výměnu dat. Rozhodnutí přejít na TEI P5 ENRICH a neukládat v metadatových záznamech hotové MARC21 záznamy zároveň s TEI 5 popisnými metadaty, snížilo podstatně v samotném Manuscriptoriu potřebu použití standardu METS jako kontejneru pro komplexní digitální dokument. Tato funkcionalita byla nahrazena standardem novým, dle [KNOLL, POLIŠENSKÝ a UHLÍŘ, 2008, s. 13] „*kvalitativně lepším*“ řešením. Standard METS se tak v Manuscriptoriu skutečně používal pouze v letech 2007-2008, nepočítáme-li analýzy probíhající od roku 2005.
Je pravdou, že pevná specifikace jednotlivých částí komplexního digitálního dokumentu je v případě agregace metadat z mnoha různých zdrojů vhodnější řešení než flexibilní METS záznam, který umožnil vložit do bibliografické části i do administrativní části v podstatě záznamy v jakémkoliv formátu. Tato různorodost pak znesnadňovala vyhledávání a následné operace nad metadaty. TEI P5 ENRICH tyto obtíže vyřešil, ale spolupracující instituce nutil provádět převody jejich interních záznamů do TEI P5 ENRICH. Sjednocení na jedné metadatové platformě bylo pro další mezinárodní spolupráci ovšem nutné.[207] Nástroje, jako např. M-Tool v nové verzi 2 online,

[206] Pozor, jen databáze, archiv zůstal v MASTER+.
[207] Standard METS by bylo možno považovat za vhodnější v případě, že by Manuscriptorium obsahovalo dokumenty jedné provenience, tedy např. pouze z NK ČR. Pak by konkrétní specifikace METS a jeho vnitřních formátů byla dobrým řešením jak pro archivaci, tak pro zpřístupnění. METS pro historické

převodníky ze standardů MASTER a EAD do ENRICH TEI P5[208], byly sice poskytnuty, ovšem volnost a jednoduchost se vytratila.

Problémem se opět ukázalo použití MARC21, ve kterém staré tisky, rukopisy a prvotisky katalogizuje několik českých knihoven. NK ČR vytváří popisy v TEI P5 ENRICH pro rukopisy a prvotisky, pouze pro staré tisky ve standardu MARC21 (ty se pak převádí do TEI P5 ENRICH). Možnost vytvářet popisná metadata pro historické dokumenty přímo v TEI P5 ENRICH a předtím ve standardu MASTER díky nástroji M-Tool vždy existovala, ne všechny knihovny v ČR ale záznamy takto vytvářet chtěly. Knihovny mají speciální báze pro rukopisy, prvotisky a staré tisky a nechtějí vytvářet záznam ke stejnému fyzickému dokumentu podruhé ve standardu TEI P5 ENRICH. Katalogizátoři z mnoha důvodů preferují vytváření záznamů ve standardu MARC21 přímo v aplikacích automatizovaného knihovního systému i přesto, že MARC21 není schopen vyjádřit všechny aspekty popisu historických dokumentů (plný text, údaje o digitálním dokumentu, strukturu dokumentu, podrobný knihovědný popis apod.). MARC21 záznam je lehké vložit do knihovního katalogu, Souborného katalogu ČR apod. Obecně lze říci, že rozpor vyplývá z potřeb různých komunit. Bibliografický popis v TEI (případně MASTER) je preferován pouze pro potřeby vědecké komunity, ne pro knihovnické účely popisu [KNOLL, 2010a, s. 26].
V souvislosti s přechodem na TEI P5 ENRICH a v návaznosti na rozsáhlou analýzu interoperability standardů TEI P5 a MARC21 [KAŠPAROVÁ a PSOHLAVEC, 2008], byl v roce 2009 obohacen nástroj M-Tool o možnost převodu z MARC21 do TEI P5 ENRICH[209] a zpět, a také o možnost z TEI P5 ENRICH vytvářet zpětně záznamy ve standardu MODS a Dublin Core [KNOLL, POLIŠENSKÝ a UHLÍŘ, 2009, s. 12]. Odpovídající XSL šablony jsou k dispozici – jejich popis viz např. [PSOHLAVEC, 2009].

Specifikace TEI 5 ENRICH našla své uplatnění ještě na dalším poli v NK ČR, a to v systému pro zpracování a prezentaci dokumentace fyzického stavu knihovních exemplářů. Aplikace systému ResIS vznikla v roce 2010 po několikaletém plánování, specifikaci metadat a funkcionality. Pro navrženou hierarchickou strukturu popisu existující a nově vznikající dokumentace restaurátorských zásahů byla uzpůsobena specifikace TEI 5 ENRICH a metadata jsou tak plně kompatibilní s TEI 5 – více o celém projektu viz [NOVOTNÝ, 2011].

5.3.1.3 Strukturovaný popis plných textů historických dokumentů

Jednou z podstatných věcí, která odlišuje digitální knihovnu historických textů, je nutnost zpřístupňování samotných textů jinak, než pouhým OCR. Používá se metajazyka, který umožňuje texty po formální a věcné stránce strukturovat. Od roku 2005 byl proto v Manuscriptoriu testován standard TEI, který umožňuje strukturovat texty i v sémiotickém smyslu [IKAROS, REDAKCE, 2005] a vytvářet tak různé edice původních textů. SGML jazyk byl na tento úkol příliš složitý, HTML dokáže popsat pouze vnější formu textu. Jako nejvhodnější se jevilo využití standardu TEI. Bylo ovšem potřeba standard pro jeho rozsáhlost a možnosti zúžit tak, aby bylo možné jeho použití [KNOLL, POLIŠENSKÝ a UHLÍŘ, 2004, s. 9]. Vznikla tak specifikace mss-fulltext, která byla v roce

dokumenty používá např. Univerzitní knihovna Heidelberg, která má velmi významné sbírky historických dokumentů.

[208] http://www.dbase.cz:8090/ege/

[209] Záznam MARC21 vytvořený podle pravidel popisu historických dokumentů v NK ČR je používán jako základ pro TEI P5 ENRICH katalogizační záznam. Ten dále může být obohacen o strukturální metadata a vzniká tak komplexní digitální dokument Manuscriptoria, pokud je to potřeba.

2005 revidována a upravena. V témže roce vznikla i definice mss-verse pro strukturaci historických plných textů ve veršované podobě (finální verze v podobě mss-verse.dtd vznikla v roce 2007). V roce 2007 spatřila světlo světa ještě definice pro strukturaci plných textů obsahujících tabulky (mss-fulltext-table.dtd). Původně všechny tyto specifikace vycházely z TEI P4 [KNOLL, POLIŠENSKÝ a UHLÍŘ, 2007, s. 10], později byly převedeny do TEI P5. Všechny definice využívají pouze minimální nutný počet polí ze všech možných tak, aby byl zápis co nejjednodušší, ale zároveň ještě dostatečně přesný.[210]

5.3.2 Metadatový popis novodobých dokumentů

Po vzniku DTD periodika a DTD monografie a zprovoznění aplikace zpřístupnění Kramerius v roce 2003 byla situace okolo metadat víceméně stabilizovaná. Proběhly pouze drobné úpravy obou DTD. V roce 2006 začala v NK ČR ve větší míře digitalizace monografií, do té doby to byla spíše periodika.

5.3.2.1 Implementace METS pro novodobé dokumenty v aplikaci Kramerius

V roce 2005 se začalo uvažovat o nasazení standardu METS pro novodobé dokumenty. První idea byla použít METS jako standard pro archivní data a metadata, která by byla uložena jako METS balíčky. METS by jako interní formát pro archivaci dat posloužil výborně, obě DTD totiž slučovala do jednoho schématu různé typy metadat – převažovala popisná, dále byla vyjádřena fyzická a později i logická struktura a v minimální míře administrativně technická metadata. Bylo by třeba ale změnit kompletně procesy výroby metadat v NK ČR i v rámci VISK7. Posléze byl standard METS použit jako výměnný standard pro sdílení metadat a dat v aplikaci zpřístupnění Kramerius, nebyl používán pro archivaci.

Obě DTD obsahují převážně popisná metadata, stejně jako metadatový popis objektů v jiných projektech u nás. METS ovšem poskytl příležitost použít i nová schémata a začít tak připravovat data z digitalizace novodobých dokumentů na procesy dlouhodobé ochrany. Bylo rozhodnuto, že do standardu METS se doplní další údaje o digitálních objektech, konkrétně administrativní a technická metadata. Vzhledem k situaci v okolním světě a v podobných institucích padla volba na standardy MIX, PREMIS a MARCXML. Nové typy metadat vznikaly (a dodnes vznikají) v digitalizaci a záznamy v nezměněné podobě jsou ukládány v archivu, kde ve složce doplňují DTD záznam metadat. DTD záznam obsahuje na PREMIS a MIX záznamy linky, takže dohromady tvoří logický archivní balíček. Na vstupu do archivu, kterým vždy byl pouze file systém, se žádná metadata, např. o událostech a změnách, nepřidávají. Implementace je tedy pouze základní.

Samotný METS záznam vzniká dynamicky (*on-the-fly*) až v aplikaci Kramerius a to na vyžádání[211]. Spustí se mechanismus, který vyextrahuje z konkrétního DTD popisná metadata, převede je do MARCXML, vytvoří METS, který obsahuje opět linky na PREMIS a MIX záznamy uložené na serveru

[210] V roce 2009 byl proveden pokus outsourcovat manuální přepis plných textů do Indie (firma Planman Technologies Inc.). Bylo přepsáno 42 knih v češtině, latině a italštině. Úspěšnost přepisu byla 90% a texty šlo bez výraznějších korektur použít [KNOLL, POLIŠENSKÝ a UHLÍŘ, 2009, s. 18]. Šlo ovšem o knihy se standardním písmem, u dokumentů s historickými písmy by bylo potřeba odborníka – paleografa nebo alespoň zběhlého historika. K ostrému využití outsourcingu nikdy nedošlo, právě pro tato omezení.

[211] METS záznam lze z aplikace Kramerius získat generováním, tedy kliknutím na tlačítko u konkrétní úrovně popisu. Další možností je získání záznamu METS automatickou cestou přes OAI-PMH, nebo jako celek (titul např.) speciální aplikací, která posbírá všechny METS záznamy jednotlivých jejích úrovní.

aplikace. Podoba METS záznamu je tak poplatná způsobu jeho tvorby. Používá převážně linkování na již hotová metadata (záznamy PREMIS a MIX) pomocí elementů <mdRef> – viz Ukázka 1. Vložený metadatový záznam přímo v METS záznamu obsahuje pouze část popisných metadat <dmdSec>, a to v podobě MARCXML a Dublin Core. METS využívá jen hotová metadata, tedy DTD s bibliografickými údaji, dále soubory PREMIS a MIX, které se začaly vytvářet od roku 2008 – viz kapitola 5.3.2.2. Žádná další metadata negeneruje ani odnikud nepřebírá. U dokumentů digitalizovaných před rokem 2008 tedy PREMIS a MIX chybí.

```
<mets:amdSec ID="AMD">
  <mets:rightsMD ID="AMD-R_METSRights">
        <mets:mdRef MDTYPE="OTHER" OTHERMDTYPE="METSRights" MIMETYPE="text/XML" L
        ABEL="MetsRights" LOCTYPE="URL"xlink:href="http://kramerius.nkp.cz/kramerius/amd?h
        dl=ABA001/11258216&amp;amp;section=rightsMD&amp;amp;type=METSRights"/>
</mets:rightsMD>
```

Ukázka 1 – Linkování pomocí <mdRef> na soubor administrativních metadat práv.

V rámci implementace proběhla analýza standardu METS a s tím spojené analýzy dalších zmíněných metadatových standardů, včetně výběru konkrétních elementů, které se měly používat. Bylo nutné také rozhodnout, jak bude vypadat finální METS záznam. Celé řešení bylo vytvářeno s přihlédnutím k jiným podobným projektům ve světě (např. univerzita v Göttingen nebo francouzský projekt *Persee*), kde se METS používal buď jako interní formát v digitálním repozitáři, nebo i pro prezentaci složených digitálních objektů. Po analýze jsme na podzim 2006 dospěli k následujícímu řešení METS záznamu. Popisná metadata ve standardu MARCXML se vkládala do části METS <dmdSec>, technická metadata PREMIS Object a MIX se vkládala do části METS <techMD>. Metadata práv našla své místo v části METS <rightsMD>. K zápisu metadat o administrativních právech k digitálnímu objektu byla zvolena část standardu PREMIS Rights, ale nikdy se neplnila. Pro popis legislativních práv sloužil standard METS Rights. METS část <digiprovMD> se využívala k zaznamenání událostí spojených s objektem a to v části standardu PREMIS Events – viz Tabulka 1. K našemu potěšení toto rozdělení bylo doporučeno jako nejvhodnější na workshopu věnovanému využívání standardu PREMIS ve Stockholmu v březnu 2007 a je podporováno i METS „redakční“ radou. Další možností je vložení kompletního PREMIS záznamu přímo pod element <amdSec>, které se ovšem příliš nevyužívá.

<techMD>	PREMIS Object
	MIX
<rightsMD>	PREMIS Rights
	METS Rights
<digiprovMD>	PREMIS Events

Tabulka 1 – Vložení jednotlivých schémat do METS záznamu, podčásti administrativních metadat <amdSec>.

Na počátku úvah o nasazení standardu METS bylo také potřeba rozhodnout o způsobu implementace METS standardu. Stávající DTD totiž popisují v jednom XML záznamu fyzický celek, který se digitalizoval. Tj. nejčastěji ročník periodika, svazek nebo více svazků monografie. Záznam ročníku periodika obsahuje popisná metadata všech čísel a příloh, včetně popisných metadat

ročníku i titulu a také logické vyjádření struktury. Jednotlivá DTD, např. ročníků, se pro aplikaci Kramerius spojují do jednoho titulu periodika. Výsledný záznam je tak dlouhý i několik tisíc řádek. Prezentace takového dokumentu v aplikaci Kramerius probíhá po jednotlivých úrovních, po kterých se uživatel pohybuje (např. u periodik je to titul, ročník, číslo, vnitřní část čísla, stránka čísla). A právě pro každou z těchto úrovní se generuje jeden METS záznam. Princip jednoho záznamu METS pro celý titul ovšem nebyl využit z obavy o velikost výsledného XML záznamu [POLIŠENSKÝ, 2007, s. 102]. Přitom standard METS je právě pro vyjádření komplikované struktury dokumentů v jednom záznamu stavěn.

DTD monografie má následující obecnou strukturu, které pak odpovídají úrovně v aplikaci zpřístupnění Kramerius: Monograph (titul), MonographUnit (svazek), MonographComponentPart (logická vnitřní část), MonographPage (stránka) – viz Obrázek 19. Pro DTD periodika je struktura: Periodical (titul), PeriodicalVolume (ročník), PeriodicalItem (číslo), PeriodicalInternalComponentPart (logická vnitřní část), PeriodicalPage (stránka) – viz Obrázek 18. METS záznamy jsou generovány pouze pro úrovně Monograph, MonographComponentPart a Periodical, PeriodicalVolume, PeriodicalItem, PeriodicalInternalComponentPart. Pro každou úroveň jeden METS záznam. Pro úroveň stránek METS nevzniká, svazek monografie je sloučen s titulem. Pro definici vazeb mezi jednotlivými METS záznamy různých úrovní je použito rozšíření standardu Dublin Core – Terms. Vazbu používá klient pro nalezení popisných a administrativních metadat vyšších úrovní. Dublin Core Terms na nižších úrovních vždy odkazují na úrovně vyšší.

Je tak možné vygenerovat METS záznam pro úroveň titul, který obsahuje popisná metadata pro titul v MARCXML a Dublin Core, odkazy na administrativní metadata a metadata práv vážící se k titulu a strukturální mapu pro všechny ročníky. METS záznam je možné vygenerovat i na úrovni ročníku periodika. Ten pak obsahuje v popisných metadatech Dublin Core jen údaj o tom, že náleží ke konkrétnímu titulu. Dále obsahuje linky na administrativní metadata a metadata práv k ročníku; přehled čísel a případně interních částí čísel a logickou strukturální mapu čísel včetně jejich dat vydání a čísla vydání – viz Ukázka 2. Poslední úroveň kde lze vygenerovat METS záznam je úroveň čísla periodika nebo pro jeho součást (např. článek). METS záznam pak obsahuje údaje o náležitosti čísla k ročníku v Dublin Core, linky na administrativní metadata čísla, a administrativní a technická metadata jednotlivých stránek čísla, pokud taková existují.[212] V logické strukturální mapě čísla nalezneme linky na jednotlivé stránky. Podobný přístup funguje i pro monografie, tam lze ovšem METS záznam vytvořit pouze pro úrovně titul a vnitřní součást (kapitola apod.).

Bibliografický popis MARCXML obsahují jen úrovně Monograph, Periodical a PeriodicalVolume. K převodu metadat do MARCXML vznikly převodní tabulky obou DTD (monografie i periodika) do MARC21.

Administrativní metadata jsou vyjádřena ve schématech PREMIS a MIX na různých úrovních popisu různě. Např. metadata práv existují pro všechny úrovně, technická metadata MIX a PREMIS Object pouze na úrovni stránky. Souborová sekce METS záznamu <fileSec> může obsahovat čtyři skupiny (fileGrp) souborů:

- URL odkazy do Krameria, na všechny úrovně (pomocí Handle identifikátorů);
- METS odkazy na soubory METS vyšších a nižších úrovní;
- IMAGE odkaz na zdroj obrazové reprezentace dané stránky;
- TXT odkaz na zdroj textové reprezentace dané stránky.

[212] U dokumentů digitalizovaných před rokem 2008 se zpětně nedoplňovala.

Podobně je koncipována i logická strukturální mapa, která vyjadřuje strukturu celé entity, tj. uvádí popis úrovně (např. ročník) a v atributech TYPE a LABEL jednotlivých elementů i popis jejích podčástí (např. PeriodicalItem) včetně data vydání, číslování. Odkaz na nižší úrovně je veden na konkrétní METS záznam – viz Ukázka 2. Atribut ORDER vyjadřuje pořadí konkrétního elementu ve strukturální mapě.

```
<mets:div ID="ABA001/1664768_SML" TYPE="PeriodicalItem" ORDER="2" LABEL="3 (17.1.1821)">
    <mets:fptr FILEID="ABA001/1664768_METS"/>
</mets:div>
```

Ukázka 2 – element logické strukturální mapy METS záznamu v aplikaci Kramerius

Fyzická strukturální mapa je využita jen na úrovni čísla (nebo výtisku monografie), které je tvořeno „fyzickými“ stránkami. Fyzická mapa čísla má atribut TYPE s hodnotou „Pages“ a obsahuje vnořené elementy pro jednotlivé soubory reprezentující jednu stránku čísla (obraz v DjVu, OCR v TXT a URL odkaz na stránku v systému Kramerius) – viz Ukázka 3. Podobně je to řešeno u svazků monografií.

```
<mets:structMap TYPE="Pages">
    <mets:div ID="SMP" TYPE="Pages">
        <mets:div ID="PG-0_SMP" TYPE="TitlePage" ORDER="0" ORDERLABEL="(1)">
            <mets:fptr FILEID="PG-0_URL"/>
            <mets:fptr FILEID="PG-0_DJVU"/>
            <mets:fptr FILEID="PG-0_TXT"/>
    </mets:div>
</mets:structMap>
```

Ukázka 3 – element fyzické strukturální mapy METS popisující titulní stránku čísla periodika.

Poslední část METS záznamu tvořeného v Krameriovi je sekce <StructLink> obsahující strukturální linky. Propojuje elementy logické strukturální mapy s elementy strukturální mapy fyzických stránek. Takto jsou pevně spojeny jednotlivé stránky díla s jeho logickou strukturou.

METS a ochranná metadata (MIX, PREMIS) byla do aplikace zpřístupnění Kramerius implementována v letech 2007 a 2008 v rámci výzkumného záměru *Optimalizace nástrojů pro digitalizaci tištěných dokumentů ohrožených degradací kyselého papíru* (2006-2010). Implementace byla uznána jako výsledek typu S – prototyp/uplatněná metodika [Implementace formátu METS v Systému Kramerius, 2007]. Na implementaci METS standardu pracovali pracovníci NK ČR, Knihovny Akademie věd ČR a technicky vše provedla firma Qbizm. Autor této práce vytvářel mapování stávajících DTD periodika a monografie do standardu MARCXML pro jednotlivé úrovně, návrh struktury METS záznamu a specifikaci polí standardu PREMIS.

5.3.2.2 Implementace PREMIS a MIX pro novodobé dokumenty v NK ČR

Přípravy na použití administrativních metadat začaly v NK ČR již v létě 2007 s souvislosti s pracemi okolo standardu METS. Specifikace ochranných metadat vycházela ze standardu PREMIS verze 1 z května 2005. Finální podoba specifikace vydaná v srpnu 2008 [HUTAŘ, 2008a] je však již uzpůsobena a upravena podle PREMIS verze 2, která vyšla v březnu 2008. Obsahovala také

technická metadata pro obrazové soubory ve standardu MIX[213]. Přináší přehled elementů, které byly vybrány k používání z jednotlivých standardů, jejich konkrétní popis i příklady použití. Pro administrativní metadata všech digitálních objektů byl použit PREMIS Object – viz Tabulka 2, pro technická metadata obrazových souborů standard MIX – viz Tabulka 3. Pro metadata práv METS Rights – viz Tabulka 4.
Celá implementace nových typů metadat byla myšlena jako pilotní příprava na budoucí využití v LTP systému, který NK ČR plánovala později nakoupit. Záměr byl takový, že bude dobré začít co nejdříve vytvářet alespoň základní technická a administrativní metadata, která se posléze využijí v LTP systému. Pro samotný LTP systém se počítalo s úpravou všech metadat, k čemuž také došlo – viz návrh profilu pro projekt NDK, který je součástí této práce. Implementací METS a nových typů metadat vznikl nový národní standard pro tvorbu metadat v digitalizaci novodobých dokumentů v programu VISK7.

Podle specifikace začala vznikat v procesu digitalizace novodobých dokumentů administrativní metadata v podobě XML souborů pro obrazové i textové soubory. Administrativní metadata pro obrazy obsahují technická metadata ve standardech MIX a PREMIS Object, dále také metadata práv v podobě METS Rights. Technická metadata pro textové soubory neobsahují MIX záznam, standard je určen pouze pro obrazová data. Bylo rozhodnuto, že nejvíce vyhovujícím přístupem bude tato nová metadata uchovávat mimo původní DTD záznam, tj. nerozšiřovat jej o nové elementy. Při změnách specifikace administrativních metadat, které byly předjímány, by bylo nutné měnit specifikaci obou DTD.
Původně se počítalo s tím, že aplikace Kramerius bude tato administrativní metadata aktivně využívat. Proto byl importní modul do systému Kramerius připraven na import i pro části standardu PREMIS Event, PREMIS Rights a PREMIS Agent. Posledně jmenované části standardu PREMIS se nikdy používat ani vytvářet nezačaly[214] a administrativní metadata v Krameriovi byla nakonec využívána pouze na dynamickou tvorbu METS záznamů pro jednotlivé úrovně digitálního dokumentu. Všechny jmenované části PREMIS jsou součástí až aplikačního profilu metadat v projektu NDK – viz kapitola 7 a příloha.
Ukázalo se tedy, že administrativní metadata nejsou pro uživatele relevantní a jsou důležitá spíše pro dlouhodobou archivaci. Administrativní metadata byla tedy pasivně uložena jako další XML soubory v balíčku spolu s daty a DTD metadaty celé entity (ročníku apod.) a tímto způsobem archivována. Tím byl naplněn primární cíl administrativních metadat, která byla od počátku myšlena jako vklad do budoucna, kdy bude potřeba dělat konkrétní operace dlouhodobé ochrany (např. migrace formátů, přesuny dat) nebo emulace na základě administrativních metadat.

[213] Specifikaci elementů provedla firma Qbizm.
[214] Z dnešního pohledu je jasné, že PREMIS Event, PREMIS Agent je nutné použít i v procesu digitalizace pro zachycení událostí jako je skenování, ořez, tvorba uživatelských kopií apod. Tehdy jsme ovšem počítali se zachycením událostí v metadatech pouze v budoucím LTP systému.

PREMIS Object

Element	Plněné hodnoty	Logická úroveň (representation)/ soubor (file)
ObjectIdentifier	n/a	Representation + File
ObjectIdentifierType	***NDK***	Representation+ File
ObjectIdentifierValue	UUID shodné s OBJID	Representation+ File
ObjectIdentifier	n/a	Representation + File
ObjectIdentifierType	Číslo zakázky	Representation + File
ObjectIdentifierValue	12345	Representation + File
PreservationLevel	n/a	Representation+ File
PreservationLevelValue	***full***	Representation+ File
ObjectCategory	***file, representation***	Representation+ File
ObjectCharacteristics	n/a	File
CompositionLevel	*0*	File
Fixity	n/a	File
MessageDigestAlgorithm	***MD5***	File
MessageDigest	MD5 hodnota otisku souboru	File
MessageDigestOriginator	IČO - 15238199	File
Size	skutečná velikost souboru	File
Format	n/a	File
FormatDesignation	n/a	File
FormatName	***JPG, DJVU, TXT, PDF, TIFF***	File
FormatVersion	***1.01 pro JPG, 25 pro DJVU, txt neplníme***	File
FormatRegistry	n/a	File
FormatRegistryName	***PRONOM***	File
FormatRegistryKey	***fmt/43*** pro JPG***, notassigned*** pro DJVU***, x-fmt/16*** pro TXT	File
CreatingApplication	n/a	Representation+ File
CreatingApplicationName	***Sirius***	Representation+ File
CreatingApplicationVersion	číslo verze	Representation+ File
DateCreatedByApplication	datum – ISO 8601	Representation+ File
Storage	n/a	File
StorageMedium	***HDD, Magnetic Tape, CD-ROM, CD-R, CD-RW, DVD-ROM, DVD-R, DVD+R, DVD-RW, DVD+RW, DVD-RAM, DVD-R DL, DVD+R DL, DVD-RW DL, DVD+RW***	File

	DL, BD-R, BD-RE	
Environment	n/a	Representation + File
Software	n/a	Representation+ File
swName	***DjVu Browser Plug-in***	Representation+ File
swVersion	***6.1.0 Build 1492***	Representation+ File
swType	***plug-in***	Representation+ File
swDependency	***Internet Explorer***	Representation+ File
Software	n/a	Representation+ File
swName	***XnView***	Representation+ File
swVersion	***1.70***	Representation+ File
swType	***renderer***	Representation+ File
Software	n/a	Representation+ File
swName	***OpenOffice***	Representation+ File
swVersion	***2.4.1***	Representation+ File
swType	***renderer***	Representation+ File
Software	n/a	Representation+ File
swName	***Windows***	Representation+ File
swVersion	***XP***	Representation+ File
swType	***operatingSystem***	Representation+ File
Hardware	n/a	Representation+ File
hwName	***128 MB RAM***	Representation+ File
hwType	***memory***	Representation+ File
hwOtherInformation	***64 MB minimum***	Representation+ File
Hardware	n/a	Representation+ File
hwName	***Intel x86***	Representation+ File
hwType	***processor***	Representation+ File
hwOtherInformation	***300 MHz minimum***	Representation+ File

Tabulka 2 – Přehled plněných elementů standardu PREMIS Object z návrhu administrativních metadat [HUTAŘ, 2008a, s. 19].

MIX

Element	**Plněné hodnoty**	**Logická úroveň (representation)/soubor (file)**
BasicImageParameters	n/a	File
Format.MIMEType	***image/jpeg, image/vnd.djvu nebo jiné***	File
File.FileSize	velikost obrázku	File
File.Checksum.ChecksumMethod	***MD5***	File
File.Checksum.ChecksumValue	MD5 souboru	File
ImageCreation	n/a	File

DeviceSource	***microfilm scanner, planetary scanner, flatbed scanner nebo jiné***	File
ScanningSystemCapture	n/a	File
ScanningSystemHardware	n/a	File
ScannerManufacturer	výrobce scanneru	File
ScannerModel	n/a	File
ScannerModelName	jméno modelu scanneru	File
ScannerModelNumber	číslo modelu scanneru	File
ScannerModelSerialNo	výrobní číslo scanneru	File
ScanningSystemSoftware	n/a	File
ScanningSoftware	jméno skenovacího software	File
ScanningSoftwareVersionNo	číslo verze software	File
ScannerCaptureSettings	n/a	File
XphysScanResolution	fyzické rozlišení scanneru (DPI) v metrech v ose X	File
YphysScanResolution	fyzické rozlišení scanneru (DPI) v metrech v ose Y	File
ImagingPerformanceAssessment	n/a	File
SpatialMetrics	n/a	File
ImageWidth	šířka obrázku v pixelech	File
ImageHeight	výška obrázku v pixelech	File

Tabulka 3 – Přehled plněných elementů standardu MIX z návrhu administrativních metadat [HUTAŘ, 2008a, s. 20].

METS Rights

RightsDeclarationMD	**Popis**	**Logická úroveň (representation)/soubor(file)**
RightsDeclaration	n/a	Representation
RightsHolder	***Zde bude umístěna deklarace práv spojená s digitálním obsahem***	Representation
RightsHolderName	***Národní knihovna ČR***	Representation
RightsHolderContact	n/a	Representation
RightsHolderContactAddress	***Klementinum 190, 110 00 Praha 1***	Representation
RightsHolderContactPhone	***420 221 663 111***	Representation
RightsHolderContactEmail	***info@nkp.cz***	Representation

Tabulka 4 – Přehled plněných elementů standardu METS Rights z návrhu administrativních metadat [HUTAŘ, 2008a, s. 21].

Kompletní popis jednotlivých elementů PREMIS Object, MIX a METS Rights s příklady plnění viz [HUTAŘ, 2008a].

Z dnešního pohledu je množina využívaných elementů velmi malá a reálné využití diskutabilní. Dnešní LTP systémy totiž tento typ metadat automaticky vytvářejí na vstupu do systému. Ukázalo se to při testování LTP systémů Rosetta a SDB v NK ČR v roce 2010, kdy se zjistilo, že z technických metadat vytvářených v NK ČR se při importu do LTP využije jen malé procento, a to většinou na zpětné kontroly (např. kontrolní součet). Údaje typu formát souboru apod. si LTP systém pomocí externích služeb jako jsou PRONOM, JHOVE vytvoří znovu sám, možnost kontroly je ale důležitá. Důležitá se ukázala administrativní metadata s údaji o výrobě. Zjistilo se, že v metadatech chybějí údaje o tom kdo a kdy skenoval, kdy a kde vznikla metadata apod. I tak byla NK ČR jednou z mála knihoven v té době (rok 2008), která vytvářela metadata ve standardu PREMIS, i když jen v jeho části Object.

5.3.2.3 Úpravy DTD monografie a periodika

V roce 2007 došlo k několika změnám ve specifikaci metadat novodobých fondů. Byly provedeny mírné úpravy stávajících DTD pro monografie a periodika (nově obě ve verzi 2.1). Byly zavedeny jednoznačné identifikátory pro každý digitální objekt odkazovaný v rámci DTD a tvořící tak komplexní digitální objekt. Došlo taky k přidání logického indexu, který vyjadřoval logické pořadí stránek v intelektuální entitě, vedle již uváděného fyzického pořadí jednotlivých stránek/souborů v rámci jedné intelektuální entity. V rámci technických metadat se začalo s přidáváním XML záznamu s kontrolním součtem pro všechny soubory obsažené v balíčku vytvořeného v digitalizaci pro konkrétní zdigitalizovaný dokument (svázaný ročník periodika, svazek monografie apod.).

Důvodem rozšíření obou DTD o tzv. index stránky byla skutečnost, že původní možnosti popisu stránky nebyly dostatečné. Neexistovala možnost jednoznačného rozlišení jednotlivých stránek, pokud např. fyzický dokument obsahoval několik kapitol číslovaných od jedné, tak se stránka s číslem 1 a dalšími opakovala [NÁRODNÍ KNIHOVNA ČR, 2007, s. 24]. Elementy <MonographPage> a <PeriodicalPage> byly proto rozšířeny o atribut Index. Přidání indexu stránky problém vyřešilo a zároveň bylo velmi jednoduché vyjádřit fyzický index – skutečné pořadí stránky ve fyzickém nebo digitálním dokumentu. Např. stránka s vytištěným číslem 5 byla reálně v dokumentu sedmá v pořadí, počítáme-li včetně nečíslovaných stránek vazby a předsádky. Z toho vyplývá, že hodnota indexu stránky musí být pro každou stránku jedinečná v kontextu jednoho digitálního dokumentu a měla by začínat od čísla 1 pro snímek první stránky digitálního dokumentu. Ukázka 4 ukazuje popis stránky monografie, která má vytištěné číslo 5, ale fyzicky je sedmá v pořadí všech stránek. Stejný princip funguje i pro periodika.

```
<MonographPage Type="Blank" index ="7">
    <PageNumber>5</PageNumber>
    <PageRepresentation>
        <PageImage href="123456.jpg"/>
    </PageRepresentation>
</MonographPage>
```

Ukázka 4 – Atribut index v popisu stránky monografie

Taktéž v roce 2007 byla stávající DTD rozšířena o index stránek u vnitřních částí monografií a periodik (MonographComponentPart a PeriodicalInternalComponentPart), kde bylo cílem zapsat rozsah stránek vnitřní části. DTD původně neumožňovala specifikovat stránky výčtem, odkazovala

pouze na první stranu vnitřní části pomocí elementu <PageNumber>, který obsahoval logické číslování, což se k tomuto účelu nehodí. Popis vnitřní části byl tedy doplněn o element <PageIndex> s atributy From a To, který může vypadat např. takto <PageIndex From="5" To="6"> – viz Ukázka 5. Využití elementu <PageIndex> umožňuje přesné určení začátku logické vnitřní části pomocí fyzického indexu stránky.

```
<PeriodicalInternalComponentPart Type="Article">
    <CoreBibliographicDescriptionPeriodical>
        . . .
    </CoreBibliographicDescriptionPeriodical>
    <PageReference>s. [3] - 4</PageReference>
    <PageNumber>[3]</PageNumber>
    <Pages>
        <PageIndex From="5" To="6"/>
    </Pages>
</PeriodicalInternalComponentPart>
```

Ukázka 5 – Element <PageIndex> a jeho využití u popisu článku v čísle periodika

DTD monografií bylo v roce 2007 také rozšířeno o elementy pro plnění unikátních identifikátorů. Rodičovský element <UniqueIdentifier> obsahoval elementy <UniqueIdentifierURNType> a <UniqueIdentifierSICIType>. DTD periodika obsahovalo SICI identifikátor již od roku 2003 od verze DTD 1.0., URN bylo přidáno ve verzi 2.0 v roce 2007. Identifikátor SICI (*Serial Item and Contribution Identifier*) je normou ANSI/NISO Z39.56-1996 (Verze 2)[215] a je určen pro jednoznačnou identifikaci části periodika, hlavně jednotlivých čísel a článků[216]. Není určen pro monografie. Rozšíření DTD monografie o tento identifikátor lze považovat za omyl. Identifikátor SICI se v NK ČR nezačal používat a element se nikdy neplnil.
Do elementu <UniqueIdentifierURNType> se později plnilo číslo UUID (*Universally Unique IDentifier*), které zajišťovalo interní jednoznačnou identifikaci jednotlivých logických entit (úrovní) digitálních dokumentů v systému Kramerius (titul, ročník, číslo aj.) jako tzv. identifikátor objektu (OBJID) – viz Ukázka 6, která obsahuje UUID úrovně stránky a také její reprezentace (souboru). UUID se využívá převážně při importu administrativních metadat do systému Kramerius, spojuje totiž dané úrovně monografie s relevantními administrativními metadaty, které vznikly jako XML soubory dodatečně a nejsou součástí DTD, musí z něj tedy být odkazovány.

```
<MonographPage Type="NormalPage" Index="9">
<MonographPage Index="0" Type="FrontCover">
<UniqueIdentifier>
<UniqueIdentifierURNType>e08f4390-19f9-11de-ac4f-000d606f5dc6</UniqueIdentifierURNType>
</UniqueIdentifier>
<PageNumber>[a]</PageNumber>
```

[215] Plný text viz http://www.pruefziffernberechnung.de/Originaldokumente/SICI.pdf.
[216] Identifikátor SICI pro číslo periodika s ISSN 0363-0277, roč. 20, č. 4, které vyšlo 23. 4. 2000, může vypadat např. následovně 0363-0277(20000423)20:4<>1.0.TX;2-V, kde „1“ za „>“ je identifikátor struktury záznamu, „0“ je označení typu popisované entity (např. abstrakt, obsah aj.), „TX“ je kód pro tištěný text, „2“ je číslo použité verze standardu SICI, „V“ je kontrolní číslo.

<PageRepresentation>
<UniqueIdentifier>
<UniqueIdentifierURNType>e1edc220-19f9-11de-a28d-00d606f5dc6</UniqueIdentifierURNType>
</UniqueIdentifier>
…

Ukázka 6 – Použití elementu <UniqueIdentifierURNType> pro plnění UUID.

Pro vytváření metadat ve standardu DTD periodika a monografie používají NK ČR, dodavatelské firmy a větší knihovny SW Sirius. Vedle toho byl od roku 2008 v Moravské zemské knihovně vytvářen open source nástroj MEditor (metadatový editor). Podnětem byla skutečnost, že aplikace Kramerius přijímá pouze již hotové metadatové záznamy odpovídající DTD specifikaci, ale na trhu neexistuje volba mezi nástroji, které by toto DTD dokázaly vytvořit. Malá knihovna, která digitalizuje jen pár dokumentů, nechce kupovat komerční SW a ani nemusí chtít zadávat tvorbu metadat externí firmě. Právě pro tyto knihovny vznikl MEditor, který umožňuje poloautomatické a uživatelsky přívětivé vytváření metadat v DTD monografie a periodika [ŠVÁSTOVÁ, 2010].

5.3.2.1 Jednoznačné identifikátory

Z důvodu zavedení permanentního a neměnného odkazu na jednotlivé úrovně digitálních dokumentů (tj. intelektuální entity jako jsou např. titul, ročník, číslo a stránka) v aplikaci zpřístupnění Kramerius byly v roce 2008 zavedeny identifikátory Handle. Identifikátor umožnil jednoznačnou identifikaci digitálních dokumentů a úrovní dokumentu při automatizovaných hromadných operacích. Identifikátor Handle v aplikaci Kramerius nevyužívá externího resolveru systému Handle[217], ale pouze jeho syntax. Handle identifikátory v Krameriovi jsou tak jedinečné pouze v konkrétní instanci systému Kramerius, ne globálně. Identifikátor Handle v Krameriovi vzniká převedením jedinečného identifikátoru UUID do syntaxe Handle. UUID verze 1 se interně používá v rámci Krameria a skládáno je z čísla síťové karty počítače na kterém vzniká, a časového razítka. Handle podoba identifikátoru je vhodnější pro uživatele, UUID totiž má 32 znaků. Odkazy na jednotlivé dokumenty jsou tedy stálé a zůstávají stejné i např. po dodatečných úpravách digitálního dokumentu v aplikaci zpřístupnění. Identifikátory lze tak použít např. při odkazování na digitální dokument v katalogu NK ČR se zárukou jejich neměnnosti.
Od roku 2011 se začíná v procesu digitalizace (projekt ANL+) používat resolver URN:NBN a přidělují se globální identifikátory URN:NBN. Ty jsou jedinečné celosvětově, ne pouze v kontextu NK ČR. URN:NBN resolver[218] byl vytvořen v roce 2010 jako pilotní provoz a počítá se s ním pro přidělování a resolvování identifikátorů pro projekt NDK.[219]

5.3.2.1 Aplikace Kramerius verze 4 a nová metadata

Už od roku 2008 se uvažovalo o přechodu od DTD periodika a monografie na standardy METS, MIX a PREMIS pro archivní data. Souviselo to s potížemi, které používání proprietárních DTD

[217] http://www.handle.net/
[218] http://resolver.nkp.cz/urnnbn/
[219] V současnosti existuje několik systémů persistentních identifikátorů určených pro digitální objekty. Ačkoliv mají rozdílné architektury, všechny jsou založeny na tomtéž principu – centrální autorita spravuje registr identifikátorů a zajišťuje získání objektu na základě zadání PID. Nejznámější jsou: DOI (správcem je DOI Foundation), ARK (California Digital Library), PURL (OCLC), Handle (CNRI) nebo URN:NBN (IANA a jednotlivé národní knihovny) [CUBR, HUTAŘ a MELICHAR, 2008, s. 1].

přinášelo v otázkách interoperability a také se závislostí na systému Kramerius. Nezanedbatelnou byla také skutečnost, že se nová verze systému Kramerius plánovala postavit nad digitálním repozitářem Fedora[220], který by umožnil ukládání a vkládání i jiných typů dokumentů, než byla periodika a monografie. Stará verze systému Kramerius by se musela upravovat a musely by se vytvořit nové specifikace DTD (např. pro vícestránkové PDF, pro audio apod.). Repozitář Fedora je schopen díky svým vlastnostem všechny tyto typy dokumentů uložit velmi jednoduše a to ve standardních formátech.

Vývoj aplikace zpřístupnění Kramerius byl od roku 2008 zajišťován v projektu VaV vedeného Knihovnou Akademie věd ČR. Ten byl zaměřen na další rozvoj aplikace a její přenesení na repozitář Fedora. Vývoj metadat tento projekt nakonec zasáhl pouze u uživatelských kopií, kde do nové verze aplikace Kramerius, verze 4 (označuje se jako K4 nebo Kramerius4) je nutné dodávat metadata ve standardu FOXML (*Fedora Object XML*). Archivní data a metadata tento projekt neovlivnil.[221] Ukázalo se, že celá změna nebude tak snadná. Bylo totiž nutno v nové verzi Krameria zachovat datový model a funkcionalitu verze předchozí. Logická struktura dokumentů zachována tedy zůstala, jen byla rozdělena na jednotlivé objekty (moduly), které je možno mezi sebou libovolně kombinovat. Nejobvyklejší kombinace odpovídají starému DTD (tj. úrovně titul, svazek, vnitřní část a stránka u monografií; obdobně u periodik). Pro články a jiné typy dokumentů bylo nutno definovat samostatný modul. Fedora repozitář opravdu umí přijímat METS dokumenty, ovšem zpřístupňovací vrstva Kramerius4 takto uložené dokumenty neumí zobrazit. Umí zobrazit pouze dokumenty ve standardu FOXML. V roce 2011 proběhla migrace starých archivních metadat z DTD do standardu FOXML, který bude pak do Fedory/Krameria4 uložen. Nové dokumenty budou vkládány také ve standardu FOXML.

Přechod na METS, MODS, MIX a PREMIS pro archivní data je tak plánován až na výrobu v rámci digitalizace projektu NDK od druhé půle roku 2012. Metadata pro uživatelské kopie pro Krameria4 pak budou z METS převáděna do FOXML. Specifikace metadat v nových standardech pro digitalizaci je podrobně popsána v kapitole 7 a příloze.

5.3.3 Metadata z pohledu dlouhodobé ochrany digitálních dat v NK ČR v období 2004-2011

Druhé období používání metadat v NK ČR charakterizují první snahy o aktivní opatření pro budoucí procesy dlouhodobé ochrany digitálních dat. První skutečné aktivity zabývající se logickou ochranou digitálních dat se tak věnovaly nevyhnutelně metadatům. V roce 2005 vzniklo první mapování mezi oběma DTD Krameria a standardy Dublin Core a MARC21. Byl to první krok k tomu, aby bylo možné převádět záznamy v DTD monografie nebo periodika do standardu MARCXML a ten vložit do záznamu METS pro intelektuální entitu odpovídající entitě popsané v původním DTD. Z analýzy, kterou provedl autor této disertační práce, vyplynulo, že možnost převodu DTD do Dublin Core je nejméně vhodná, protože Dublin Core není díky malému počtu polí schopno pojmout valnou většinu údajů z DTD. Problémy nastanou i při převodu do MARC21, ovšem v daleko menší míře. MARC21 je robustní standard, ovšem není příliš vhodný pro popis

[220] http://fedora-commons.org/

[221] Ve stejném projektu vznikl i dnešní portál Registr digitalizace, pod původním názvem RELIEF III. Slouží jako správní systém k evidenci digitalizovaných a mikrofilmovaných dokumentů. Obsahuje identifikátory, údaje o digitalizaci, vlastnících, stavu digitalizace apod. Lze říci, že obsahuje administrativní a některá technická metadata, která v samotných metadatech u digitálních objektů vznikajících v digitalizaci chybějí.

digitalizovaných dokumentů.[222] Bibliografické údaje lze přenést bez větších problémů, technické údaje již ne. Ve stejné době vznikl také základní přehledový dokument popisující kontejnerový standard METS a jeho principy – viz [HUTAŘ, 2005].
V roce 2007/2008 byly zavedeny do procesu digitalizace novodobých dokumentů standardy PREMIS, MIX, což významně přispělo výhledu na budoucí logickou ochranu metadat. Šlo o první použití těchto standardů cílené na ochranu dat, navíc poté povinné pro instituce účastnící se programu VISK7. Bohužel se to týkalo pouze digitalizace novodobých dokumentů. V obou projektech (*Kramerius* a *Manuscriptorium*) se začal využívat standard METS, v *Manuscriptoriu* pouze přechodně. Jak moc jsou administrativní a technická metadata podstatná se ukázalo v roce 2010 při testování LTP systémů, které tato metadata vyžadují, aby měly ucelenou informaci o životním cyklu digitálního objektu, od okamžiku jeho vzniku (digitalizace).
V roce 2005 vznikla také první pravidla zápisu polí pro standardy DTD periodika a monografie. Bylo to díky poučení z migrací metadat v roce 2003, kdy se ukázalo, že do konkrétních polí plní ručně pracovníci různé obsahy a ani způsob zápisu konkrétního pole ve více záznamech není konzistentní. Tato skutečnost samozřejmě působila a působí problémy při jakémkoliv automatizovaném procesu, nejen při migracích, ale i při uživatelském vyhledávání a při správě dokumentů administrátorem.[223] Konzistence obsahů jednotlivých polí metadat je důležitá pro případ, kdy jsou metadata uložena v databázi (např. LTP systému) a administrátor potřebuje vybrat konkrétní množinu záznamů na základě obsahu konkrétního pole v těchto záznamech. Zvláště důležitá jsou pravidla v případě, že metadatové záznamy v konkrétní specifikaci produkuje několik institucí a takto vytvořené záznamy jsou určeny pro uložení a zpřístupnění v jedné centrální instituci, jak je tomu např. u výstupů z programu VISK7. V roce 2006 byla připravena i pravidla pro zápis metadat monografií pro projekt *Kramerius*, doplnila tak předchozí pravidla pro periodika.

Z pohledu kontroly integrity digitálních objektů je velmi důležité, že od roku 2007 byla data vytvářená v projektu *Kramerius* (VISK7) standardně obohacována o kontrolní součty (MD5), ty jsou pak součástí archivního balíčku. Kontrolní součty jsou z hlediska dlouhodobé ochrany digitálních dat užitečné ve více ohledech. Při jakémkoliv přesouvání dat i metadat je možné automaticky a hromadně zkontrolovat integritu souborů před a po přesunu, právě porovnáním kontrolních součtů – původního a nově vytvořeného. Pokud se se souborem nic nestalo, nedošlo k žádné ztrátě nebo změně bitstreamu, jsou kontrolní součty shodné. Tento princip lze využít i na pravidelné kontroly uložených dat v úložišti, tj. kontrola archivu. Podmínkou ovšem je, aby první kontrolní součet byl vygenerován na zcela jistě bezchybných souborech a aby tento součet byl součástí buď balíčku dat, nebo byl zapsán v metadatech ke konkrétnímu objektu. Ideální je obojí.
Data vytvořená do roku 2007 kontrolní součty neměla, musely být vytvářeny zpětně při okamžiku uložení do archivu. Docházelo tak k situacím, kdy vznikl kontrolní součet pro soubor pro dokument, který se během přesunu z digitalizace do archivu poškodil.[224] K digitalizovaným

[222] Tento nedostatek odstranil formát MODS, který byl vytvořen přímo pro popis digitálních dokumentů a z formátu MARC21 vycházel.
[223] I s danými pravidly se však ukázalo, že pokud jsou pole metadatového záznamu plněna ručně lidmi, žádná pravidla nepomohou a obsahy i jejich forma se mohou lišit. Řešením je např. automatické přebírání bibliografických záznamů z katalogu instituce a tedy automatické plnění metadatového záznamu, které v NK ČR u novodobých dokumentů probíhá od roku 2008.
[224] Takový kontrolní součet je nepoužitelný, poškozený soubor se při kontrole jeví jako by byl v pořádku, kontrola nehlásí žádnou změnu v bitstreamu, který je od počátku poškozen.

historickým dokumentům z projektu *Manuscriptorium* začala firma AiP Beroun dodávat kontrolní součty po naléhání ze strany Odboru digitální ochrany a IT NK ČR bohužel až v roce 2010. Po jednáních o nutnosti a výhodách kontrolních součtů mezi AiP Beroun, oddělením IT NK ČR a Odborem digitální ochrany NK ČR byl požadavek na výrobu kontrolních součtů akceptován. Všechna data přijatá z digitalizace historických dokumentů do centrálního digitálního repozitáře do té doby dostávala kontrolní součet při procesu uložení.

Rok 2009 a následující byly v NK ČR ve znamení přípravy projektu NDK, zaváděly se do procesů nástroje na extrakci metadat (JHOVE, DROID/PRONOM), další nástroje se testovaly (PLATO). Tyto nástroje významně pomohly tvorbě a validaci metadat. Musely být zabudovány do stávajících procesů, protože funkční LTP systém neměla NK ČR k dispozici.

6. Vývoj konceptu správy a uložení digitálních dat

Z předcházejících kapitol lze vypozorovat proměny způsobu nazírání na uložení digitálních dat. Zpočátku se digitální data ukládala na fyzický nosič, který byl v době vzniku digitálních objektů aktuálně dostupný (magnetické pásky, optické disky, HDD v podobě diskových polí, páskové knihovny). Pro knihovny byly nové technologie zároveň přínosem i problémem. Úložná média jsou nestálá a od začátku bylo jasné, že jde o velký problém, který nemá srovnání v papírovém světě starých knihoven – viz např. [LESK, 1992]. Správa takto uložených dat byla velmi omezená, pohled na ni se ovšem vyvíjel, zvláště od okamžiku, kdy množství dat začalo být takřka nezvládnutelné. Bylo potřeba najít způsob, jak data efektivně pomocí konkrétních systémů spravovat a udržovat. Uložení dat a metadat na jakémkoliv nosiči a odpovídající politika zálohování se nazývá tzv. *bitstream preservation*, tedy základní dlouhodobá *ochrana bitstreamu*. Tento proces probíhá zcela běžně ve většině institucí, které ukládají data, která jsou pro ně důležitá. Rozšířeným omylem ovšem je záměna nebo sloučení ochrany bitstreamu s logickou ochranou. Logická ochrana digitálních dat je proces, kterým zajišťujeme použitelnost, vyhledatelnost a srozumitelnost konkrétních digitálních dat v blízké nebo daleké budoucnosti. Zálohy mohou za určitých okolností zajistit dostupnost digitálních dokumentů v nezměněné podobě v budoucnu, nezajistí ovšem jejich použitelnost a srozumitelnost.

Vývoj uložení dat v paměťových institucích je ilustrován na příkladu NK ČR, který zahrnuje všechny nejdůležitější technologické změny a koncepty. Uložení většího množství dat začalo v NK ČR probíhat v souvislosti s digitalizací v polovině 90. let 20. století.

6.1 Optické disky

Partnerem NK ČR v digitalizaci historických fondů byla od počátku firma AiP Beroun. Právě díky digitalizaci historických fondů musela NK ČR ve spolupráci se zmíněnou firmou začít řešit otázky vhodného uložení vyprodukovaných dat. Vznikl „digitální archiv“ optických disků CD-R. Na konci 90. let šlo o finančně nejdostupnější technologii uložení velkého množství dat [PSOHLAVEC, 2004]. Celý archiv sestával ze systému několika kopií jednotlivých dat. Optické disky byly ve dvou kopiích na dvou různých lokalitách. Archiv CD-R byl používán nejen pro výstupy digitalizace historických fondů (program *Memoria Mundi Series Bohemica*), ale i pro uložení CD-R přicházejících do NK ČR v rámci povinného výtisku. V letech 2000-2001 probíhal v rámci výzkumného projektu VaV *Optimalizace archivace a zpřístupnění digitálních dat* výzkum stárnutí a uložení optických médií. *„Otázkou bylo, nakolik lze stanovit určitý algoritmus redukované kontroly uchovávaných digitálních dat na základě statistického vyhodnocení měření CD.“* [KNOLL a PSOHLAVEC, 2002, s. 2] Měření chybovosti disků proběhlo, za sledované období bylo zjištěno minimum chyb (BLER). Na základě zjištěných hodnot probíhala kontrola stavu archivu na základě statisticky vybraných vzorků [KNOLL a PSOHLAVEC, 2002, s. 4 a 13] i nadále, protože negativní vliv času na optické disky, byť je jakkoliv pomalý, se zastavit nedá.

Je důležité si uvědomit, že v případě archivu CD-R disků z programu *Memoriae Mundi Series Bohemiae* byly pozitivní výsledky měření dány i skutečností, že se jednalo o výjimečný archiv s vhodně nastavenými pravidly a klimatickými podmínkami uložení. Archivní optické disky byly před zápisem pečlivě vybírány, kontrolovány, probíhalo měření vzorků jednotlivých šarží i měření

kvality zápisu CD-ROM mechanik. Využití těchto disků bylo minimální. Druhá kopie disků, která byla určena jako uživatelská, měla pravidla kvality o něco méně striktní. Za těchto podmínek lze konstatovat, že archiv CD disků byl na svou dobu odpovídajícím řešením schopným za nastavených podmínek zajistit dlouhodobé uchování digitálních dat.[225] Problémem se ukázalo, že od počátku nového tisíciletí, s masovým rozšířením optických disků a jejich zlevněním, začalo docházet ke snižování kvality disků i u renomovaných výrobců.

Od počátku nového tisíciletí bylo jasné, že optické CD/DVD disky nejsou vhodné médium pro dlouhodobé uložení dat ani pro jejich dlouhodobou ochranu – srovnej např. [BYERS, 2003, s. 2] nebo základní dokument popisující hrozby spojené s ukládáním dat na vypalované optické disky, který vznikl v rámci programu UNESCO *Paměť světa* [BRADLEY, 2006]. Částečně díky citlivosti a nestálosti materiálů, ze kterých jsou vyrobeny, jako jsou např. barviva, slitiny kovů (stříbro, zlato, hliník apod.). Podstatnější hrozbou je však zastarávání SW i HW. V dnešní době jsme svědky toho, že mnohé notebooky nemají CD-ROM mechaniky, podobně jako např. tablety. Velmi blízko je doba, kdy mechaniky nebudou dostupné ani na stolních PC a disky nebude kde přehrát. Optické disky jsou dnes nahrazovány jinými typy pamětí, např. flash paměti. Optické disky ne zcela dobře umožňují správu a údržbu dat, neustálé sledování integrity. I to je jeden z důvodů, proč nejsou vhodným médiem pro potřeby dlouhodobé ochrany digitálních dat. Riziko poruch a ztráty dat je u archivů založených na technologii optických disků vyšší, než u jiných dostupných technologií a přístupů [BRADLEY, 2006, s. 4]. I z těchto důvodů optické disky pro potřeby archivace velkých objemů dat uvolnily, (nejen) v paměťových institucích, místo ve prospěch integrovaných datových repozitářů, které: *„Jsou považovány za nejvhodnější pro dlouhodobé uložení i dlouhodobou ochranu dat. Kontrolují automaticky např. integritu dat, data obnovují, mohou je přemísťovat a to vše s minimálním zásahem lidského faktoru. Pokud se nad tuto HW vrstvu posadí vhodný systém na správu repozitáře, umožní přímou správu dat i metadat …*“ [BRADLEY, 2006, s. 4]

I přesto archiv optických disků se zdigitalizovanými historickými dokumenty přetrvává v NK ČR až dosud. Optické disky s daty z digitalizace historických fondů jsou stále produkovány jako počátkem tisíciletí a jsou ukládány v Odboru historických a hudebních fondů NK ČR. Probíhá i jejich měření. Zároveň je kopie těchto disků zasílána online od zpracovatelské firmy AiP Beroun přímo na datový repozitář NK ČR, a to teprve od roku 2010 [VRBICKÝ, 2011]. Roční přírůstek tohoto CD v archivu je okolo 400 kusů. Stávající archiv CD disků s daty z digitalizace historických fondů byl taktéž v roce 2010 převeden a uložen na Centrální datové úložiště a zálohován na magnetických páskách.

Optické disky přicházející do NK ČR v rámci povinného výtisku jsou také stále archivovány ve fyzickém archivu. Roční přírůstek na konci první dekády 21. století byl okolo 4000 CD/DVD disků. Celkový počet optických disků získaných v rámci povinného výtisku byl v roce 2010 31.000 kusů, z nichž zhruba polovina je hudebních disků, druhá polovina jsou datové disky [KNOLL, 2010b, s. 32]. Obsah těchto optických disků je velmi různorodý – audio, video, přílohy periodik se SW apod.

Již v roce 2006 se pro větší bezpečnost plánoval převod obsahu disků z povinného výtisku na Centrální datové úložiště NK ČR [KNOLL, POLIŠENSKÝ a UHLÍŘ, 2006, s. 11]. Při testování migrace obsahu tří set optických disků byly použity open source nástroje a uložení do ISO obrazu disku. ISO obraz disku ve formátu file systému ISO 9660 by měl umožnit, že různé operační systémy mohou použít tento ISO soubor a přistupovat k němu jako k fyzickému disku pomocí virtuální mechaniky.

[225] Srovnání poskytuje měření CD disků z povinného výtisku, které proběhlo v následujícím roce. Chybovost byla daleko větší, vzhledem ke stáří i opotřebovanosti jednotlivých disků. U 43% bylo zjištěno poškrábání, u 15% výrobní vada disku atp. – viz [KNOLL a PSOHLAVEC, 2002, s. 5].

Různé operační systémy tak uvidí soubory na disku, není ale řešena otázka jejich čitelnosti v odpovídajícím SW. Uložení ISO image řeší jen zálohování dat, tj. dlouhodobou ochranu bitstreamu. Neřeší logickou dlouhodobou ochranu dat. Pro zpřístupnění obsahů datových disků v budoucnu se počítá s budoucí emulací, protože obsahují velmi různorodá data – dokumenty, hry, SW, java aplikace, flash animace apod. Pro obsahy hudebních a audiovizuálních disků se počítá s migrací – převodem audio stop do podoby souboru za uchování určité kvality a jejich postupné migrace v budoucnu.

Testování převodu obsahu disků bylo prvním krokem k převodu všech optických disků z povinného výtisku na Centrální datové úložiště. Rutinní převod probíhá od roku 2007, je ovšem často z nedostatku financí a někdy také kvůli nedostatku úložného prostoru na Centrálním datovém úložišti přerušován. Převod logicky začal u nejstarších médií z roku 1996. V roce 2010 vznikl první návrh na nové workflow převodu disků na ISO image, které zahrnovalo např. vytváření kontrolních součtů a popisných i technických metadat, validace apod. Tyto procesy ve stávajícím workflow převodu optických disků na ISO image naprosto chybí a vlastní převod tedy nelze nijak kontrolovat a ISO image nejsou připravené na logickou dlouhodobou ochranu, nemají základní metadata – více viz [HUTAŘ a MELICHAR, 2010].

6.2 Magnetické pásky

Magnetické pásky jsou využívány na dlouhodobé uložení, ať již online nebo offline. Používají se v páskových knihovnách, kde lze jednotlivé pásky (*cartridge*) vyměňovat podobně jako magnetické disky. V NK ČR bylo vzhledem k nárůstu množství dat a snižování kvality optických disků rozhodnuto o nákupu nového řešení uložení dat, než byly optické disky. Rozhodnutí koupit robotickou magnetopáskovou knihovnu se opíralo: „ … *o stav rozvoje v ukládání objemných dat v letech 1998-1999, kdy kapacita pásek vysoko převyšovala pevné disky a rovněž jejich spolehlivost z hlediska archivace dat byla v té době nejvyšší. Robotická knihovna a související zařízení (například diskové pole a různé systémy obslužných komponent, ať operační systémy nebo databázové či jiné aplikace) vytvořily velmi komplexní celek, který byl náročný na obsluhu (OS Solaris a Linux; ORACLE databáze, SAM FS, AIP SAFE a postupně další menší subsystémy).*“ [KNOLL, 2004, s. 4] Páskový robot byl pořízen v roce 1999 v rámci projektu VaV *Digitalizace mikromédií* a byl používán jak pro uložení dat, tak pro jejich zpřístupnění. Data z digitalizace (archivní i uživatelské kopie) byla ukládána na páskovém robotu, vždy na třech páskách, 2 online, třetí kopie pásky byla uložena offline v jiné budově. Pro historické dokumenty byla čtvrtou kopií kopie v archivu CD disků [KNOLL, 2000, s. 7].

Celý systém uložení dat fungoval na SAM FS (*Storage Archive Manager File System*) s aplikační vrstvou AIP SAFE. SAM FS dokázal zajistit automatické kompletní obnovování dokumentů mj. tím, že kontroloval expirační lhůty úložných médií (pásky AIT2 a AIT3) a data z těch, které se blížily konci své životnosti, bezpečně přenášel (tj. dekomprimoval, dopočítával v případě ztrát, komprimoval) na nová média, vždy ve dvou kopiích na dvě média. Správu dat prováděla aplikace AIP SAFE, dovolovala vytváření, uložení a zpřístupnění digitálních dokumentů. Vytváření dat a metadat a také zpřístupnění pomocí indexace metadat bylo prováděno na discích, které byly součástí celého systému. Metadata uložená na discích měla zálohu pouze databázovou a v rámci zálohy samotných disků. Proto se přistoupilo i k zálohování metadat na CD-ROM disky offline [KNOLL, POLIŠENSKÝ a UHLÍŘ, 2004, s. 19].

Popsaný systém měl určité problémy v oblasti zpřístupnění, např. rychlost odezvy při potřebě vykopírovat data z pásek a poskytnout uživateli; při zpracování a tvorbě metadat; problémy s tabulkami v relační databázi; chybějící kontrola konzistence dat (!); malá bezpečnost proti útoku zvenčí, napájení apod. [KNOLL, POLIŠENSKÝ a UHLÍŘ, 2004, s. 30]. Bylo proto rozhodnuto o oddělení výrobní, archivní a zpřístupňující vrstvy, tj. oddělení archivních a uživatelských dat. Aby toho bylo dosaženo, byl od roku 2003 vyvíjen systém Kramerius[226]s GNU GPL licencí na zpřístupnění v lokálních sítích, který byl určen pouze na zpřístupnění. Kramerius umožnil poskytování přístupu daleko lepším způsobem, než byla problematická distribuce CD-ROM disků s digitálními dokumenty do jednotlivých knihoven, jak ji popsal PhDr. Jiří Polišenský: „*Dokumenty jsou rozděleny často na mnoho částí, fyzická média jsou náchylná na poškození, rozdělení dokumentu podle velikosti médií a záznam dat jsou časově náročné operace apod. Knihovny proto již dříve požadovaly mít možnost zpřístupnit digitalizované dokumenty prostřednictvím lokální sítě a CD-R média uložit jako operativní zálohu.*“ [POLIŠENSKÝ, 2004]
Na uložení dat byl nadále používán magnetopáskový robot (využíván jako jedna za záloh až do roku 2009/2010[227]) a na výrobu dat i metadat pak nově SW Sirius. Archiv dat byl tedy pouze file systém na páskovém robotu, bez dalšího SW pro správu těchto dat. Vrstva AIP SAFE nepodporovala nové standardy DTD periodika a monografie, nebyla dále plněna a aplikace nebyla dále využívána.

S rozvojem diskových úložných systémů to v půli první dekády 21. století vypadalo, že páskové knihovny budou nahrazeny úložnými systémy diskovými. Celý vývoj ovšem nakonec směřoval k tomu, že disková pole jsou většinou používána na uložení dat, která je potřeba mít rychle dostupná (např. uživatelské kopie a metadata) a archivní data, ke kterým se přistupuje méně nebo vůbec, jsou uložena v tzv. páskových knihovnách, např. s technologií LTO. Nevýhodou magnetických pásek zůstává delší dostupnost dat, díky sekvenčnímu přistupovaní k datům. Výhodou ovšem je kapacita, která se v dnešní době rovná hodnotě 3TB na jednu pásku, a také energetická nenáročnost. Podobným vývojem prošla i NK ČR a k diskovému datovému úložišti (viz kapitola 6.3.1) provozovala také páskovou knihovnu IBM umístěnou v Centrálním depozitáři Hostivař. K té v roce 2009 přibyla další páskovou knihovna, umístěna do hlavní budovy v Klementinu. Od roku 2009 jsou tak všechna data v NK ČR archivována na dvě pásky LTO4 WORM na dvou geograficky vzdálených lokalitách. Ve stejném roce byly servery NAS zapojeny do GPFS clusteru pro vyšší dostupnost dat [KNOLL, POLIŠENSKÝ a UHLÍŘ, 2009, s. 5].

6.3 Digitální repozitáře, DAM a LTP systémy

Nejužívanější pro střednědobé skladování dat jsou v dnešní době magnetické disky. Hodí se na zpřístupnění, jsou rychlé v odezvě na požadavek. V polovině prvního desetiletí 21. století došlo k nárůstu kapacit pevných disků a zároveň k poklesu ceny za 1MB úložného prostoru. Na trhu se objevily systémy na uložení dat v podobě diskových polí, někdy diskových úložišť (SATA disky). Velké objemy dat se začaly ukládat na disková pole, na magnetické pásky nebo na kombinaci

[226] Dalším důvodem vzniku aplikace Kramerius byla i potřeba rychlého zpřístupnění digitálních kopií poškozených dokumentů po povodních v roce 2002.
[227] Díky nesprávné údržbě systému ze strany dodavatelských firem i díky nadměrnému prodlužování jeho životnosti došlo v roce 2009 ke zjištění, že některá uložená data z dostupných pásek nelze vykopírovat ven a NK tak o část svých dat uložených na robotu přišla.

obého. Diskuze o řešení uložení velkého množství dat pro NK ČR začaly již roku 2004. Diskuze vyústila do rozhodnutí vybudovat Centrální datové úložiště, které mělo poskytnout prostor pro data z projektů *Manuscriptorium*, *Kramerius* a *WebArchiv*, včetně archivu CD disků. Objem dat u jednotlivých projektů byl již takový, že nemělo smysl ukládat je odděleně, ale naopak centrálně v datovém úložišti. Nově plánované úložiště také mělo být základem navrhované České digitální knihovny, jak ji popisuje *Koncepce trvalého uchování knihovních sbírek tradičních a elektronických dokumentů v knihovnách ČR do roku 2010* [Koncepce trvalého uchování…, 2005].

Studie posuzující nároky a potřeby NK ČR byla vyhotovena v roce 2005, včetně návrhu textu pro výběrové řízení. Bylo navrženo řešení postavené na dvou diskových systémech, uložených na dvou lokacích (Klementinum a Hostivař), které by obsahovaly shodná data, a prováděly mezi sebou replikaci. Tento přístup byl posléze v následujících letech i implementován.[228] V rámci příprav na nákup datového úložiště byla vypracována analýza, která formulovala mj. následující body [KNOLL, POLIŠENSKÝ a UHLÍŘ, 2005, s. 12]:

- stávající technologie bude nutno neustále migrovat na nové verze a generace;
- zařízení a média pro archivaci dat musí mít výrobcem garantovanou životnost minimálně 10 let[229];
- je lepší, když v jedné lokalitě je použito zařízení jiného výrobce a jiného stáří, než v lokalitách ostatních;
- k archivním datům není nutný rychlý přístup, který by zvýšil cenu jejich uložení;
- zařízení na archivaci dat musí být co nejvíce otevřené, použitelné na více platformách a neproprietární, nevázané na služby nebo technologie jednoho dodavatele;
- náklady na pořízení tvoří pouze zlomek celkových nákladů na HW, na jeho provoz a vlastnictví – řešením může být outsourcing uložení dat externím firmám.

Tyto zásady jsou platné dodnes. Outsourcing je častou volbou především v západních národních knihovnách, v NK ČR není pro uložení dat využíván. Často, podobně jako v projektu NDK, je nutno pořídit HW jako investice a ne jako službu. Přitom outsourcing uložení velkého množství dat může být v řadě případů a ohledech výhodnější.[230]

Datový repozitář NK ČR byl nakoupen v roce 2006 a plně implementován v roce 2007 (firma T-Systems jako dodavatel). Šlo o systém IBM DS4800, který byl na dvou lokalitách, v Klementinu a zrcadlený v Hostivaři. Ve stejném roce byly zahájeny přesuny dat.

Ovšem ani datové úložiště nemusí být zárukou odpovídající správy dokumentů. Dokáže ukládat data do file systému, má základní nástroje na zálohování (např. TSM/HSM). Pokud ale není k dispozici systém na správu dat s připojenou databází, který by uchovával informace o integritě dat; o lokaci souborů konkrétního dokumentu; o změnách; o přemístění apod., je to velké ohrožení celé snahy o ochranu dat. V případě NK ČR bohužel tato vrstva nad úložištěm chybí. Správu dat nahrazuje excelová tabulka, která obsahuje lokace jednotlivých dokumentů, jejich metadat, obsahuje identifikátory (číslo zakázky aj.), podle kterých lze najít, o jaký konkrétní

[228] Až po nějaké době provozu obou systémů se ukázalo, že politika „okamžité“ replikace dat nebyla nejvhodnější. Došlo totiž i k replikaci chyb, tj. pokud na jedné lokaci došlo k havárii se ztrátou dat, tato ztráta nebo poškození se replikovalo i na druhou lokaci a NK ČR tímto způsobem přišla o určité objemy archivních dat.

[229] V současnosti je běžná obměna HW po zhruba pěti letech, morální zastarání je někdy dokonce rychlejší.

[230] Poskytovatelská firma má odborníky na danou problematiku, přebere náklady na údržbu, na výměnu a nákup nových generací zařízení, je flexibilnější v případě nárůstů objemů dat nebo naopak, apod.

dokument se jedná. Digitalizované historické dokumenty a jejich metadata v projektu *Manuscriptorium* mají od roku 2007 vlastní systém správy dat, postavený na systému pro správu dokumentů AIP SAFE III. Jedná se ovšem o archiv zpracovatelské firmy AiP Beroun. Tento systém nepracuje nad Centrálním datovým úložištěm NK ČR. Správa archivních kopií historických dokumentů v NK ČR je zajišťována také excelovou tabulkou.
NK ČR tedy v současné době nedisponuje žádným systémem umožňujícím správu archivních digitálních dokumentů. Nikdy nebyl nad daty z digitalizace zaveden žádný z dostupných DAM systémů na správu repozitáře (*Digital Asset Management*), jako je např. DSpace, DigiTool aj. V úložištích, která nejsou pouhým file systémem, ale jsou spravována k tomu určeným DAM systémem, se metadata na vstupu do repozitáře importují do databáze systému, aby bylo možné s nimi pracovat (upravovat, nahrazovat, měnit, doplňovat, vyhledávat apod.). DAM systémy pracují s metadaty konkrétních lokálně nebo vzdáleně uložených digitálních objektů. Správa dat v archivu NK ČR byla po celou dobu odkázána pouze na náhradní řešení. To znamená, že NK ČR nemá k dispozici žádný efektivní nástroj pro úpravu nebo obohacování metadat v úložišti. Dokumenty jsou na páskách „zamrzlé“, ve stavu, v jakém byly před 10 lety uloženy, ačkoli by bylo žádoucí obohacovat archivní balíčky o nové informace. Zatímco například data v digitální knihovně Kramerius procházejí neustálým zpracováním, jsou upravovány jejich struktury a metadata, dokumenty a metadata v archivu tyto změny nereflektují. Současný způsob archivace dat poskytuje správcům jen velmi omezené možnosti analýzy obsahu repozitáře či kontroly kvality dat a metadat, nemluvě o získávání informací pro vykazování finanční náročnosti archivace dat z jednotlivých projektů [FOJTŮ, HUTAŘ a MELICHAR, 2011, s. 75]. Z tohoto stručného popisu je zřejmé, jakým ohromným rizikům jsou data v úložišti NK ČR vystavena. Zcela chybějí nástroje pro efektivní správu a dlouhodobou ochranu dokumentů v úložišti. Několikanásobné zálohování dat na médiích různého typu nelze považovat za dlouhodobou ochranu digitálních dokumentů.
Zachování bitstreamu, který dokumenty tvoří, je základním předpokladem jejich budoucí logické dlouhodobé ochrany, ale k uchování intelektuálního obsahu dokumentů a zajištění jejich dlouhodobé použitelnosti to samo o sobě nestačí. K tomu je potřeba repozitář v souladu s požadavky normy OAIS a nástroji jako TRAC nebo *Nestor Catalogue of Criteria for Trusted Digital Repositories* [FOJTŮ, HUTAŘ a MELICHAR, 2011, s. 76]. Odpovídající kvalitu uložení ve všech potřebných, i netechnických směrech, je možné zajistit podle dvou konceptů, které tuto problematiku řeší a definují konkrétní náležitosti a zásady. Prvním z nich je referenční rámec OAIS, který v popisu modulů *Archivní sklad*, *Správa dat* a *Administrace* specifikuje náležité procesy vedoucí k dlouhodobému a spolehlivému uložení a uchování digitálních objektů. Jde např. o proces obnovy médií, přenosy dat, kontrolu integrity bitstreamu, záchranu dat v případě havárií, předcházení zastarání HW (vše modul *Archivní sklad*). Druhým konceptem je tzv. důvěryhodný repozitář, jehož funkce, procesy a nutné vlastnosti jsou také definovány – více viz kapitola 6.4 a deset zásad důvěryhodného repozitáře. Právě do konceptu důvěryhodného repozitáře se promítl referenční rámec OAIS velmi důkladně.

Nutno říci, že vyspělé paměťové instituce mají již fázi budování digitálních repozitářů s odpovídajícím SW na správu dat víceméně zvládnutou, digitální, a někdy i důvěryhodné, repozitáře mají a bez obtíží je provozují. Existují komerční i open source DAM systémy, které jsou často nasazovány (DSpace, DigiTool, CONTENTdm, Fedora aj.). Od konce prvního desetiletí 21. století jsou ovšem tyto instituce zaměřeny na další krok a tím je logická dlouhodobá ochrana digitálních dat v podobě LTP systému. Staví tak na zkušenostech se správou dat a někdy také

přímo na již provozovaném repozitáři. Při nasazení LTP systému se veškeré světové knihovny orientují na prověřená a otevřená řešení. Jdou cestou sdílení znalostí a specifikací požadavků na nové systémy, popisů datových modelů a metadat. „*Zkušenost s provozem první generace systémů pro dlouhodobou ochranu je totiž naučila, že jedním ze zásadních požadavků na systém pro dlouhodobou ochranu a správu digitálních dokumentů je otevřenost ve smyslu možnosti integrace nástrojů třetích stran, veřejné dokumentace, flexibility nastavení data modelu a jednotlivých workflow pro správu dat.*“ [FOJTŮ, HUTAŘ a MELICHAR, 2011, s. 74] Funkcionalita LTP systémů překračuje běžnou správu dat a ochranu ve smyslu záloh. Nabízejí logickou ochranu, sledování a přidávání metadat, automatizované funkce na vyhodnocování rizik, nástroje na migrace, na plánování opatření ochrany, případně mezinárodní komunitu uživatelů.
Ostatní, v tomto smyslu rozvojové státy, stále zápasí s tím, aby měly odpovídající digitální repozitář a problematika dlouhodobé ochrany pro ně není zatím aktuální. NK ČR se víceméně rozhodla přeskočit fázi s nasazením DAM systému a v rámci projektu NDK stanovila jako jeden ze tří hlavních cílů nákup a nasazení LTP systému. Celý projekt včetně přesunu archivních dat byl plánován tři roky. Nyní (konec roku 2011) to vypadá, že dodané řešení bude systém na správu firemních digitálních dokumentů (ECM – *Enterprise Content Management*), do kterého se firma bude snažit dodat LTP funkcionalitu. Toto řešení a výběr firmy ohrožují projekt NDK samotný. Zdá se tedy, že NK ČR etapu DAM systému nepřeskočí.

V současnosti existuje mnoho platforem a systémů na tvorbu a provozování digitálních repozitářů, ale jen málo z nich se soustředí na logickou dlouhodobou ochranu digitálních objektů, jak ji specifikuje referenční rámec OAIS. Často chybí procesy plánování ochrany apod. Pro oblast správy záznamů vznikla i sada mezinárodních norem ISO 23081, z nichž nejvíce relevantní je první část z roku 2006 *Information and documentation -- Records management processes -- Metadata for records -- Part 1: Principles*[231]. Ta je postavena na další obecnější normě ISO 15489 *Records Management standard*[232].

V mnoha institucích se také zkoumá využití uložení dat v tzv. *cloudu*, který by se dal využít jak pro uložení, zálohování, tak pro logickou dlouhodobou ochranu. Je to vlastně další médium, se kterým se SW pro repozitáře nebo LTP učí pracovat.[233] Nejčastěji jsou využívány služby Amazon S3, Microsoft a další. Populární je také SW iRODS, který zajišťuje správu digitálních objektů uložených v *cloudu*. Na jeho základě funguje např. *data grid* rozhraní (a také síť) pro dlouhodobou ochranu digitálních dat pocházejících z více institucí – Chronopolis[234], vyvíjený v SDSC (*San Diego Supercomputer Center*), iniciovaná Kongresovou knihovnou v rámci NDIIPP programu.

6.3.1 Definice digitálního repozitáře

Digitální repozitář je systém sestávající z HW a SW, který má za cíl získávat, uchovávat (krátkodobě nebo i dlouhodobě), spravovat a zpřístupňovat digitální informace, které jsou do něj ukládány, ať již automaticky nebo ručně. S obecnou definicí digitálního repozitáře to není tak jednoznačné, repozitářů existuje několik druhů, které se liší nejen svou politikou, ale hlavně se liší a závisí na

[231] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=40832
[232] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=31908
[233] LTP systém SDB od firmy Tessella uložení do cloudu umí, stejně jako systém Rosetta od firmy Ex Libris.
[234] https://chronopolis.sdsc.edu/

potřebách organizace, která je provozuje. Takovou organizací může být univerzita, knihovna, archiv, výzkumný ústav apod. Pro ilustraci, cílem určitého repozitáře může být uchovávat informace a poskytovat je ve srozumitelné podobě uživatelům, zatímco cílem jiného repozitáře je také uchovávat informace, ke kterým ale poskytuje přístup pouze administrátorům. Repozitáře paměťových institucí jsou ve valné většině určeny pro zprostředkování přístupu k jejich obsahu pro běžného uživatele. Jako obecná definice by se tedy dala použít tato: "*Místo, na které lze uložit různé digitální materiály, a které má přidaný smysl v tom, co ukládá a k jakému účelu je provozováno.*" [CARPENTER, 2005]
Z uvedeného je jasné, že repozitář není pouze technické řešení, nejde pouze o sbírku digitálních objektů, ale jde také o doprovodné záležitosti okolo HW a SW. Repozitář a správa v něm uložených dat závisejí na kvalitě metadat. Ve zkratce by měl každý repozitář, resp. DAM SW nad ním běžící, provádět následující činnosti, dle [REESE, 2008, s. 56-58]: Měl by přijímat jakékoliv formáty, měl by být škálovatelný, bezpečný. Mělo by být možné kontrolovat stav dat a úložných médií, včetně reportů o metadatech a systémové aktivitě. Repozitář musí ukládat popisná, strukturální, administrativní metadata tak, aby bylo možno je doplňovat a schéma upravovat. Repozitář by měl automaticky zaznamenávat do metadat a do databáze časy vstupu a změny jednotlivých objektů, včetně přidělování interních identifikátorů systému. Uložená data nesmí být závislá na konkrétním HW a ani na DAM systému samotném. Repozitář by měl umožnit přístup k datům, měl by případně tvořit uživatelské kopie archivních dat, měl by umět zpřístupnit pouze metadata, nebo jejich část. Zároveň musí udržovat integritu archivních dat. Měl by mít standardní rozhranní, např. OAI-PMH, SRU/SRW apod. Data uložená v repozitáři musí být samozřejmě zálohována. Nejčastěji jsou data uložena na discích a další kopie ještě na páskách v dalších lokalitách. Disky jsou vhodné pro zpřístupnění, data jsou rychle dostupná, z pásek je dostupnost pomalejší.

Výraz digitální repozitář se mnohdy zaměňuje za digitální knihovnu. Digitální repozitář je přitom nejčastěji pouze částí digitální knihovny, pokud za digitální knihovnu považujeme celek skládající se z repozitáře, HW a aplikace pro zpřístupnění. Hlavním cílem repozitáře je uložení a správa obsahu, zatímco hlavním smyslem digitálních knihoven je zpřístupnění případně sdílení digitálních dokumentů, jak může vyplývat i z této široké definice DLF (*Digital Library Federation*), která definovala pro své potřeby digitální knihovnu jako: „ *… organizaci poskytující zdroje včetně odborných pracovníků k výběru, pořádání, zpřístupnění, interpretaci, distribuci, ochraně integrity a odolnosti vůči času pro digitální zdroje tak, že tyto jsou ihned k dispozici k využití definovanou komunitou.*" [WATERS, 1998] Definice digitální knihovny se podobá definici digitálního repozitáře, důraz je ovšem kladen na zpracování a zpřístupnění. Zároveň obsahuje i funkcionality, které jsou z dnešního pohledu typické pro repozitáře nebo LTP systémy (udržení integrity dat, odolnost vůči času), autor tedy chápal digitální knihovnu jako digitální repozitář[235].

Digitální repozitáře jsou logickým rozšířením služeb klasických knihoven. Koncept budování repozitáře je ovšem velmi odlišný od klasických dokumentů. Budování digitálního repozitáře často znamená změny organizační struktury a zaběhnutých procesů nebo vytvoření procesů zcela nových. Platí to zvláště pro metadata, kde musí být k dispozici odborník, který s metadaty je

[235] V roce 1998 se ještě digitální repozitář, digitální knihovna a systémy na dlouhodobou ochranu digitálních dat příliš nerozlišovaly jeden od druhého. Jednotlivé aspekty problematiky digital preservation však již byly známy – viz [GARRETT a WATERS, 1996].

schopen pracovat, posuzovat jednotlivá schémata a jejich vhodnost pro konkrétní projekty, musí být schopen metadata vytvářet, upravovat, včetně kompletních specifikací jednotlivých schémat.
Vytvářet digitální repozitář není jednorázová činnost. Jde o nekončící závazek, který není hotov s pořízením HW a naistalováním SW na správu dat. Jde o neustálý rozvoj, financování, správu, školení pracovníků. HW je nutno obměňovat, data je potřeba migrovat na stále nové a nové formáty, což stojí vše finanční prostředky. Na tuto skutečnost upozorňují snad všechny práce věnující se digitálním repozitářům, příkladem může být [REESE, 2008, s. 5]. Digitální repozitáře jsou závislé na technologiích a může lehce dojít k podcenění nákladů na správu. Zpočátku, když je vše nové, HW i SW má podporu, náklady budou zcela jistě nižší než za pár let, kdy např. firma dodávající SW již nebude existovat, odejdou pracovníci a noví nebudou mít přehled o jednotlivých funkčních částech repozitáře, ani o tom, jak byly vyvíjeny. V takovém případě se náklady na údržbu rapidně zvyšují.

6.3.2 DAM (Digital Asset Management) systémy

Většina institucí ukládajících velká množství dat na svém úložišti provozuje nad tímto úložištěm nějaký DAM systém. DAM systémy jsou určeny primárně na správu dat a metadat, jejich procesní zpracování, uložení, vyhledávání. Dalšími výhodami použití takového systému jsou mj.:

- využití sítě pro lepší přístupnost dat než je třeba na optických discích,
- podpora zavedených standardů,
- podpora komplexních objektů,
- možnost vytváření virtuálních sbírek,
- podpora základních opatření ochrany digitálních dat (např. sledování a záznam údajů o změnách dat i metadat),
- centrální vyhledávání a přístup k datům,
- verzování dokumentů i metadat,
- zamezení duplicitnímu uložení shodných objektů,
- správa a sledování přístupů, autorizací a opatření bezpečnosti.

Je možné rozlišit typy SW pro správu dat podle časového úseku, pro který data ukládají. Některé se zaměřují primárně na zpřístupnění dat, které spravují pouze krátkodobě (např. uživatelské kopie budou v dohledné době nahrazeny) a nemají tedy propracovanou správu ani procesy pro dlouhodobou ochranu. Některé mohou ukládat digitální objekty ve střednědobém úseku (např. deset let), další v dlouhodobém úseku (více než deset let). Do první skupiny náleží aplikace zpřístupnění (např. Kramerius), do druhé skupiny by bylo možné zařadit právě DAM systémy, do poslední skupiny specializované LTP systémy.
Ze zástupců nejznámějších open source DAM systémů uveďme např. systémy DSpace[236], Fedora[237], CONTENTdm[238], Greenstone[239], EPrints[240], Invenio[241]. Z komerčních produktů je běžný např. DigiTool[242] od firmy Ex Libris. Fedora a DSpace jsou rozšířeny hodně v USA, EPrints ve Velké

[236] http://www.dspace.org/
[237] http://fedora-commons.org/
[238] http://www.contentdm.org/
[239] http://www.greenstone.org/
[240] http://www.eprints.org/
[241] http://invenio-software.org/
[242] http://www.exlibrisgroup.com/category/DigiToolOverview

Británii v prostředí univerzit. Tvůrci systémů Fedora a Eprints se snaží dodávat do systémů funkcionalitu jdoucí za prostou správu digitálních objektů a přidávají funkcionality podporující dlouhodobou logickou ochranu digitálních dat.

- ***Fedora*** uvádí jako svou výhodu schopnost ukládat a spravovat všechny typy digitálních objektů spolu s metadaty a škálovatelnost pro několik milionů objektů. Potenciál Fedory objevily záhy americké univerzity, které ji využívají k uložení dat i jejich zpřístupnění. Fedora podporuje většinu známých standardů metadat, jako METS, RDF na vstupu do systému. Vlastní metadata jsou uložena v kontejnerovém standardu FOXML (*Fedora Object XML*). Data jsou oddělena od ostatních částí systému, je proto možná jejich migrace na jiný repozitář. Architektura odpovídá OAIS.
- ***DSpace*** SW byl vyvíjen s vidinou dlouhodobé ochrany digitálních dat. Mnohé procesy dlouhodobé ochrany podporuje (kontrolní součty, kontrola integrity, autenticity, podpora otevřených formátů metadat apod.), ale není na stejné úrovni jako specializované LTP systémy, což platí i pro Fedoru a EPrints. DSpace není příliš flexibilní, umožňuje pouze „plochý" a jednoduchý model metadatových schémat, omezena je i hierarchie digitálních objektů.
- ***EPrints*** nebylo zamýšleno jako SW na dlouhodobou ochranu, ale pro správu a zpřístupnění digitálních dokumentů a akademických prací. V posledních letech je ovšem funkcionalita pro dlouhodobou ochranu doplňována, zvláště ve spolupráci s projektem PLANETS v letech 2008-2010. Ukazuje se, že je možné do existujících DAM systémů doplnit funkcionality logické dlouhodobé ochrany. To bylo i cílem britského projektu KeepIt[243], který z tohoto pohledu porovnával čtyři SW. Projekt se snažil o integraci nástrojů a služeb do existujících systémů. TO se povedlo zvláště u SW EPrints, do kterého byly implementovány funkce podporující identifikaci formátů (DROID), charakterizaci formátů (XCDL, XCEL), podpora verzí, analýza rizik spojených s formáty (PRONOM, PLATO) a možnost použít plán ochrany vytvořený v nástroji PLATO, který automaticky provede zamýšlené opatření, včetně kontroly kvality. První takto obohacenou verzí EPrints byla verze 3.0 z roku 2007.

Za systém na správu a ochranu dat lze považovat také open source SW LOCKSS[244] (*Lots of Copies Keep Stuff Safe*). LOCKSS je aplikace, vyvinutá v roce 1999 na univerzitě ve Stanfordu, která se nainstaluje do běžného PC. V něm si ukrojí určitou část pevného disku a na ni ukládá data. Jednotlivá PC se potom spojují v síť a tvoří tak vlastně distribuovaný systém uložení více kopií dat. LOCKSS je klasickou ukázkou řešení pasivní ochrany, kde jeden digitální objekt je uložen v rámci sítě na více lokalitách. V případě, že se objekt v jedné lokaci začne lišit od ostatních svých kopií na dalších lokacích (např. poškození), ostatní lokality, které jej také ukládají, rozhodnou, která podoba je správná. Využití nástrojů jako jsou JHOVE, DROID na identifikace a charakterizaci formátů je teprve uvažováno jako předpoklad k pozdějším aktivitám logické dlouhodobé ochrany digitálních dat [SCHULTZ a GORE, 2010, s. 109]. LOCKSS sítě v různých obměnách a variacích

[243] http://www.jisc.ac.uk/whatwedo/programmes/inf11/digpres/keepit.aspx
[244] http://www.lockss.org/lockss/Home

vznikají po celém světě, nejaktivněji v USA. Známá je privátní síť *MetaArchive*[245]. Dalším zástupcem jsou privátní sítě PeDALS (*Persistent Digital Archives and Library System*) aj.
Aplikace LOCKSS odpovídá referenčnímu rámci OAIS [LOCKSS ALLIANCE, [2008]]. V ČR systém nikdo v ostrém provozu nepoužívá. V Německu je na LOCKSS postavený projekt LuKII (*LOCKSS und KOPAL Infrastruktur und Interoperabilität*)[246] na ochranu digitálních dokumentů (2010-2012). Podle slov Tobiase Steinkeho ovšem toto nasazení není odpovídajícím řešením dlouhodobé ochrany digitálních dat, které hledá a chce využívat německá národní knihovna. Toto řešení je pouze základní a může pomoci menším knihovnám ve sdílení nákladů na dlouhodobou ochranu bitstreamu. K požadované funkčnosti bude nutno vytvořit pro LOCKSS novou funkcionalitu [STEINKE, 2011].

6.3.3 Přehled dostupných LTP systémů

V posledních pěti letech se začala objevovat jak hotová komerční řešení, tak také pokusy o open source LTP systémy. Pro organizace velikosti NK ČR je vhodným kandidátem spíše komerční řešení, které nevyžaduje velké úpravy a je schopno prokazatelně pracovat s miliony digitálních objektů. V komerčním řešení, které se již někde používá, je jistá záruka stability a úspěšné implementace. Toto open source nástroje ve většině nabídnout nemohou, hodí se proto pro střední nebo malé instituce, případně pro malé projekty, které mají zdroje na vývoj, úpravy a další rozvoj systému.
Z komerčních byly v předchozích kapitolách zmíněny Safety Deposit Box[247] (SDB, výrobce firma Tessella, Velká Británie); Rosetta[248] (výrobce firma Ex Libris, Izrael). Donedávna byl komerčně dostupný také DIAS[249] (výrobce firma IBM), který ale v oblasti knihoven nemá další vývoj.
S open source LTP systémy se v posledních letech setkáváme stále častěji. Jsou odpovědí na vysokou cenu komerčních systémů a také výstupem často několikaletého vývoje, který se začal zúročovat až v poslední době. Velkou množinu zástupců tvoří systémy postavené na SW pro správu repozitáře Fedora (RODA, HOPPLA, MOPSEUS, ISLANDORA). Další jsou výstupem vývoje buď knihoven, archivů nebo specializovaných komunit. Níže jsou některé ze systémů popsány.

- ***DAITSS***[250] je SW pro dlouhodobou ochranu digitálních dat vyvíjený ve *Florida Center for Library Automation* (FCLA). Ve verzi 1 fungoval od roku 2005, ve verzi druhé, kompletně přepracované, od roku 2011 – více viz [CAPLAN a CHOU, 2011]. Je zamýšlen jako tzv. dark archiv, tj. nemá modul zpřístupnění pro uživatele, lze jej ovšem použít v součinnosti s nějakou aplikací zpřístupnění. Odpovídá OAIS, využívá METS a PREMIS a jeho cílem je provádět aktivní (logickou) ochranu digitálních objektů formou migrace, k čemuž využívá dostupných a volných nástrojů a metadat.
- ***RODA***[251] je open source SW vyvíjený od roku 2006 Národní archivem Portugalska ve spolupráci s Univerzitou Minho tamtéž. Jde o jeden z nejvyspělejších LTP systémů, které

[245] MetaArchive je projekt hrazený z NDIIPP (*National Digital Information Infrastructure and Preservation Program*), který vzešel z prostředí univerzit v USA, které využívají tuto síť na replikace, distribuci a uložení digitalizovaných dokumentů na více lokacích (tj. pasivní ochrana bitstreamu).
[246] http://www.d-nb.de/wir/projekte/lukii.htm
[247] http://www.digital-preservation.com/solution/safety-deposit-box
[248] http://www.exlibrisgroup.com/category/RosettaOverview
[249] http://www-935.ibm.com/services/nl/dias/
[250] http://daitss.fcla.edu/
[251] http://redmine.keep.pt/projects/roda-public?locale=en#home

jsou volně dostupné. RODA podporuje existující standardy, jako jsou OAIS, METS, PREMIS, EAD a je postavena na repozitáři Fedora. Systém má uživatelská rozhranní pro všechny moduly OAIS, administrátor tedy může sledovat a nastavovat procesy jako vyhledávání, administraci uložených dat, metadat, provádět procesy vkládání dat a ochrany. Ochranná metadata jsou generována automaticky v PREMIS schématu. RODA ve spolupráci s aplikací CRiB (viz kapitol 3.2.5.3) podporuje opatření logické dlouhodobé ochrany, jako jsou: identifikace datových formátů, návrhy na optimální formátové migrace, vlastní migrace, kontrola výsledků migrace, generování ochranných metadat. RODA se dále rozvíjí v rámci SCAPE projektu.

- ***ARCHIVEMATICA***[252] je v současné době spolu se systémem RODA nejrozvinutější volně dostupný LTP systém. Je vyvíjen v Kanadě komunitou odborníků a ve spolupráci s archivy města Vancouveru a Mezinárodního měnového fondu. Projekt je podporován výborem *Paměť světa* UNESCO. Jedná se od počátku o LTP systém na logickou ochranu dat určený pro instituce střední velikosti. Řešení systému spočívá ve spojení volně dostupných tzv. mikroslužeb do jednoho integrovaného SW řešení (balíku) spolu s workflow postaveném na upravené distribuci Linux Xubuntu. Mikroslužby pro *digital curation* spolupracují s běžným file systémem. Výhodou je, že každá služba/nástroj může být zaměněna za další, ať již novou verzi nebo kompletně nový nástroj. Mikroslužby/nástroje jsou shodné jako v ostatních LTP systémech, které je ovšem používají jako plug-iny. Archivematica využívá některých vlastností open source aplikací, jako je např. Inkscape, OpenOffice apod. Systém odpovídá OAIS, má grafická rozhraní podobně jako RODA na většinu procesů a nastavení. Archivní balíčky obsahují automaticky generovaná metadata ve schématech METS, PREMIS a Dublin Core. Systém obsahuje modul plánování ochrany dle OAIS. Více viz [VAN GARDEREN, 2010].
- ***MOPSEUS*** [253]– systém vyvíjený v Řecku jako management systém pro digitální knihovnu s funkcionalitou logické dlouhodobé ochrany digitálních dat, včetně implementace plánů ochrany. Aplikace je postavena nad repozitářem Fedora, který poskytuje správu služeb v kombinaci se základními workflow a funkcionalitou. MOPSEUS poskytuje hotový a k použití připravený systém pro repozitář, na rozdíl od Fedory, která potřebuje četné úpravy směrem ke kustomizaci. Cílem je poskytnou jednoduchý nástroj na uložení a logickou ochranu dat pro malé a střední instituce, pro které je použití velkých řešení jako je RODA, Archivematica nebo komerčních systémů náročné a někdy i nemožné. Více k systému viz [GAVRILIS, PAPATHEODOROU, CONSTANTOPOULOS a ANGELIS, 2010]. Systém se spoléhá na vnitřní formát Fedory, tj. FOXML specifikaci metadat, ale lze definovat své schéma. Datový model systému je mapován na PREMIS datový model, je tedy možné metadata ve standardu PREMIS automaticky na vstupu a při změnách objektů generovat a ukládat v repozitáři Fedora v rámci FOXML metadat.
- ***HOPPLA***[254] – LTP systém vyvíjený na technické univerzitě ve Vídni (TUW). Je určen pro malé instituce nebo domácnosti. Cílem HOPPLA systému bylo vytvořit základní a na ovládání jednoduchý, co nejvíce automatizovaný systém pro logickou ochranu v domácím využití. HOPPLA kombinuje zálohy bitstreamu a migrační workflow pro migrace datových

[252] http://archivematica.org/
[253] http://194.177.192.14/mopseus/index.html
[254] http://www.ifs.tuwien.ac.at/dp/hoppla/

formátů ohrožených zastaráváním. Samozřejmě spolupracuje s externími službami a nástroji třetích stran, jako je např. DROID na identifikaci formátů. Využíván je také JHOVE pro automatické generování metadat. Informace o tom, kdy je nutno konkrétní formát migrovat je založena na úsudku expertů, kteří toto rozhodnutí distribuují z centrálního místa do všech napojených HOPPLA systémů, které jsou online (samozřejmě na vyžádání). Samotná migrace probíhá lokálně v systému konkrétního uživatele a to nástroji, které má v rámci systému HOPPLA nebo ve svém počítači. O těchto nástrojích ví centrální služba, která údaje o PC a systému HOPPLA sbírá z tzv. profilu, který si každý uživatel systému vytvoří. Systém samotný obsahuje mnohé open source knihovny k migracím volně využitelné, např. Java OpenDocument pro migrace DOC2PDF, DOC2ODT, PPT2PDF atd. – více viz [STRODL, et al., 2010].

- ***EPrints***[255] – původně DAM systém na správu akademických prací vyvíjený na univerzitě v Southamptonu. Je ukázkou toho, že lze omezeně funkcionalitu LTP do DAM systémů doplnit. Systém je od roku 2009 schopen pracovat s výstupy nástroje na plánování ochrany PLATO a provádět na jeho základě migrace dat, zobrazovat verze apod. EPrints je tak doplněn o modul plánování ochrany a odpovídá více OAIS referenčnímu rámci. Stává se z něj jednoduchý LTP systém.

6.3.4 Rozdíly mezi LTP a DAM systémy

Hlavním cílem DAM systému je poskytnout prostředí pro manipulaci, správu a zpřístupnění dat. Zatímco hlavním cílem LTP systému je obsah dlouhodobě ochraňovat. DAM systémy často obsahují zárodky logické dlouhodobé ochrany, ale z pohledu OAIS jim chybí podstatné části, které by takový systém opravdu dělaly schopným logické ochrany. Uložení a správa digitálních objektů, kterou provádějí, neřeší problém zastarávání formátů. Při vkládání digitálních objektů do DAM systému jsou administrativní a technická metadata přidávána v omezené míře, na rozdíl od LTP systému, kde je toto klíčové. Na vstupu do DAM systému neprobíhá validace ani identifikace formátů tak, jak ji známe z OAIS a z LTP systémů.

DAM systémy často nemají za cíl ukládat data dlouhodobě. Postrádají proto OAIS modul *Plánování ochrany*, nepočítají příliš s tím, že by se v rámci systému prováděla například formátová migrace za účelem dlouhodobé ochrany. Právě absence logické dlouhodobé ochrany, tj. kombinace modulů *Plánování ochrany* a *Administrace* (OAIS) v podobě procesů jako je identifikace formátů, validace formátů, automatická tvorba administrativních, technických a ochranných metadat je hlavní rozdílem mezi DAM a LTP systémem. Modul *Plánování ochrany* poskytuje sledování vývoje technologií (díky registrům formátů a risků s nimi spojených), které pak dodává administrativnímu modulu podnět k ochranné aktivitě založené na předem připraveném ochranném plánu. Modul *Plánování ochrany* umožňuje takový plán vytvořit, upravovat, testovat apod.

6.3.5 Digitální repozitář a moduly referenčního rámce OAIS

Již několikrát jsme zmínili, že repozitář má odpovídat OAIS referenčnímu rámci (*OAIS compliant*). Co to ovšem reálně znamená? Výraz „odpovídá OAIS“ nemá žádné oficiální pozadí v podobě instituce, která by takové hodnocení udělovala. Neexistuje ani žádná oficiální sada pravidel/podmínek, při jejichž splnění by bylo možné prohlásit, že digitální repozitář je OAIS

[255] http://www.eprints.org/

kompatibilní. Jediným měřítkem je referenční rámec OAIS sám. Na celou problematiku existují minimálně tři pohledy [OCLC/RLG PREMIS WORKING GROUP, 2004, s. 26-27]. První pohled říká, že kompatibilita je popsána přímo v referenčním rámci OAIS, v částech 1.4, 2.2 a 3.1. Příkladem takového pojetí může být prohlášení kompatibility systému LOCKSS s OAIS[256], které popisuje, jak LOCKSS odpovídá nárokům z částí 2.2 a 3.1 OAIS referenčního rámce. Druhý pohled je takový, že digitální repozitář odpovídá OAIS tehdy, pokud repozitář má funkcionalitu takovou, jak je specifikována v OAIS nebo alespoň implementuje hlavní charakteristiky modelu OAIS, jako je šest klíčových funkčních komponent a specifikované odpovědnosti. Třetí pohled je v podstatě odmítavý a říká, že kompatibilita OAIS je velmi vágní termín a referenční rámec není konceptem, ke kterému by měla být jakákoliv kompatibilita vztahována.
Jak je tedy patrné, skutečnost, že konkrétní SW nebo digitální repozitář proklamuje, že odpovídá OAIS, neznamená, že má plnou funkcionalitu z OAIS vyplývající a tím méně to znamená, že se jedná o LTP systém. Kompatibilita OAIS není rovna funkcionalitám LTP systému. Ve velké většině digitálních repozitářů chybí klíčový modul *Plánování ochrany* a s ním spojené procesy.

Popis a funkcionalita modulů archivního systému, jak je specifikuje referenční rámec OAIS (a kapitola 3.3.2), je níže přiblížena na příkladu specifikace funkcionality LTP systému [MELICHAR a HUTAŘ, 2011, s. 41-48] a jeho modulů pro projekt NDK, který z OAIS vychází. Vznik textu specifikace byl významně ovlivněn testováním dostupných komerčních LTP systémů v roce 2010. Tehdy proběhlo v NK ČR v Odboru dlouhodobé ochrany digitálních dat na evropské poměry zcela výjimečné POC (*Proof of Concept*) se systémy SDB (Tessella) a Rosetta (Ex Libris). Oba nástroje mají workflow zpracování vstupu dat založená na pravidlech a využívají podobné služby i nástroje (PRONOM/DROID, JHOVE, NZME) pro validaci balíčků, formátů dat, extrakci technických metadat. Oba systémy umožňují efektivní práci s daty v archivu – doplňování metadat a nových reprezentací digitálních objektů jednotlivě i hromadně, přidávání identifikátorů, přeskupování dokumentů či změnu jejich struktury. Oba systémy mají též integrované nástroje pro identifikaci formátových rizik, modul plánování ochrany, který umožňuje testovat a provádět například migrace formátů. Oba systémy udržují všechna metadata o životním cyklu dokumentu v archivu podle OAIS.
Nákup hotového komerčního LTP systému byl jeden ze tří hlavních cílů NDK. Měl přinést řešení logické dlouhodobé ochrany stávajících i nově vzniklých digitálních objektů, která je v NK ČR zanedbána. Systém má uchovat informační obsah (digitální objekty) a veškerá potřebná metadata pro procesy logické ochrany a zajistit tak použitelnost výstupů digitalizace a digital-born dokumentů v budoucnu.

6.3.5.1 Modul Příjem (Ingest)

První fáze vstupu do digitálního repozitáře je zajištěna modulem *Příjmu* (*ingest*). V této fázi by měly přicházet již balíčky SIP, připravené k uložení, tedy ve formátu LTP systému. Modul provádí především automatické identifikace a validace formátů souborů nebo validaci a kontrolu technických metadat, které již s balíčkem přijdou od producenta, tzv. *enrichment* (obohacení metadat automatickou cestou), případnou migraci do preferovaných formátů (normalizace), opětnou antivirovou kontrolu. Toto bude probíhat zapojením služeb třetích stran jako je JHOVE, PRONOM/DROID, NZME a další. Průběh pohybu vstupního balíčku musí být vidět v online

[256] http://www.lockss.org/lockss/OAIS

monitorovacím modulu systému, kde bude informace dostupná správci repozitáře i dodavateli dat a tato metadata se budou dále archivovat. Průběh balíčku musí být řízen pravidly konfigurovatelným workflow, složeným z kroků, které je možné do workflow vkládat individuálním nastavením nebo automaticky na základě uložených pravidel. Tj. konkrétní způsob zpracování se musí řídit tím, od jakého producenta, jaký typ dat je vkládán. Systém musí umět automaticky nastavit pravidla pro řešení chyb nástrojů pro validaci a charakterizaci. Musí při příjmu umět na základě interní knihovny formátů identifikovat formátová rizika a automaticky migrovat do bezpečných formátů. Výstupem z modulu *Příjem* je balíček AIP připravený k archivaci.

6.3.5.2 Modul Archivní sklad (Archival Storage)

V archivním modulu dochází k uložení již zpracovaných dat v podobě balíčků AIP. *Archivní sklad* je nejpodstatnější částí celého systému a má vazby na všechny ostatní moduly. Základem je vlastní technologie uchovávání digitálních objektů na fyzických úložištích. Modul obdrží a ukládá AIP balíčky z modulu *Příjem* a zajišťuje správu těchto balíčků. Systém musí mj. umožnit nastavení pravidel pro uložení různého typu materiálu různým způsobem tj. na fyzická úložiště různého typu a do různých struktur nebo logických částí. Zajišťuje také komunikaci s *middlewarem* (HSM apod.) a se systémy přímo obsluhujícími HW úložiště. Musí udržovat informace o lokacích jednotlivých digitálních objektů tak, aby bylo možné je vždy získat. Mazání archivních položek (intelektuálních entit) nebo části položek (např. metadatové záznamy) by mělo být možné provést pomocí tzv. soft a/nebo hard-delete[257]. Velmi důležité je, že modul musí udržovat informace o všech verzích/reprezentacích dat nebo metadat (po migracích apod.), případně jen některé vybrané verze dat/metadat. V případě selhání musí být k dispozici kontrolní součty pro hlídání integrity a způsoby obnovení dat. Způsob, jakým jsou archivní balíčky uloženy na HW úložiště, musí být zdokumentovaný a dostupný. Modul *Archivní sklad* poskytuje data/metadata pro modul *Zpřístupnění* v podobě DIP balíčku.

6.3.5.3 Modul Administrace (Administration)

Modul *Administrace* zahrnuje služby a funkce podporující správu LTP systému a jeho nastavení, správu vztahů a smluv s producenty/dodavateli dat, nastavení standardů pro vstup do archivu, správu konfigurací HW a SW systému, správu účtů správců aplikace a případných uživatelů a také autentikaci a reporting. LTP systém musí mít propracovaný modul pro monitoring různého typu: sledování pohybu objektů v jednotlivých fázích životního cyklu balíčků; sledování a zaznamenávání všech prováděných akcí, včetně monitoringu použitých formátů a SW; podpora monitoringu distribuce dokumentů z repozitáře, sledování historie transakcí a operací prováděných uživateli a dodavateli dat; kontrola změn systémové konfigurace – možnost uložit a zálohovat nastavení celého LTP systému. Modul by měl také udržovat knihovny formátů a s nimi svázaných dokumentů a odpovídajících aplikací k jejich zobrazení. Důležitá je také správa účtů a profilů producentů/dodavatelů dat. Účty obsahují údaje a pravidla pro standardy dat a metadat, která producenti poskytují, včetně povolených objemů dat, struktury balíčků, maximální velikosti souborů, povinných polí metadat, lokace dat, informace o workflow, které se použije na vstup do

[257] Soft-delete vymaže údaje o digitálním objektu z databáze, objekt samotný na úložišti zůstává. Hard-delete funkcionalita vymaže objekt zcela.

repozitáře apod. Modul *Administrace* spojuje ostatní funkce a rozhraní s producenty, uživateli a modulem *Správa dat*.

6.3.5.4 Modul Správa dat (Data Management)

Modul *Správa dat* poskytuje služby a funkcionalitu k vytváření, udržování, vyhledávání a zpřístupnění metadat, která popisují a identifikují archivní dokumenty, včetně administrativních metadat sloužících k řízení archivu (archivního modulu). Funkce modulu zahrnují administrování databázových funkcí (údržba schémat, definicí), updaty databáze (nové popisné informace apod.), provádění dotazů na data z archivního modulu, poskytování odpovědí na dotazy i vytváření reportů. Modul musí umožňovat nalezení metadat souvisejících s konkrétním digitálním objektem, tento objekt pak na základě specifických požadavků musí jít stáhnout, exportovat, upravit. Musí být možné vytvořit nebo změnit metadata k digitálnímu objektu, přičemž změny a updaty jsou uchovávány v metadatech samotných.

6.3.5.5 Modul Plánování dlouhodobé ochrany (Preservation Planning)

Modul *Plánování ochrany* je pro LTP systém klíčový a je tím, co LTP systém odlišuje od DAM systémů, které tento modul postrádají. LTP systém monitoruje základní vlastnosti vkládaného materiálu již v modulu *Příjem* a na základě získaných metadat inteligentně pomáhá správcům repozitáře s plánováním dlouhodobé ochrany (musí držet informace o vložených formátech a platformách, na kterých fungují; o použitých metodách komprese a dalších souvisejících technologiích, které mohou mít potenciálně dopad na použitelnost archivovaného materiálu). Modul musí být dostatečně otevřený, aby mohl využívat i nově vzniklé nástroje na plánování ochrany a akce ochrany v budoucnu (nástroj PLATO apod.). LTP systém musí udržovat lokální registr formátů a s nimi souvisejících aplikací, včetně jejich závislostí, ideálně s možností sdílení s ostatními uživateli shodného LTP systému. Modul musí umožňovat:

- automatický nebo manuální výběr množiny objektů z archivu dle konfigurovatelného dotazu, tedy na základě vlastností digitálních objektů;
- výběr strategie dlouhodobé ochrany za použití služeb třetích stran;
- vytvoření plánu na dlouhodobou ochranu;
- testování možností a alternativ vybraného plánu dlouhodobé ochrany;
- porovnání výsledků testů různých opatření – automatické nebo manuální;
- implementaci plánu na dlouhodobou ochranu vytvořených v nástrojích mimo LTP systém;
- provedení vybraného způsobu ochrany;
- validace testů i finálních výsledků;
- monitorování cílové komunity a vývoje technologií.

6.3.5.6 Modul Zpřístupnění (Access)

Modul musí umožnit vyhledávání a dodání archivovaných digitálních objektů a jejich metadat v různě strukturovaných balíčcích DIP. Odpověď na uživatelskou žádost o přístup k digitálnímu obsahu může obsahovat výsledný balíček DIP obsahující pouze metadata spojená s příslušnými entitami; nebo pouze jeden či více digitálních objektů bez metadat; případně také kombinaci obého, dat i metadat. Modul *Zpřístupnění* musí být také schopný řešit situace, kdy není možné

digitální objekt z archivu získat, tedy informovat, že objekt není dostupný, uživatel nemá oprávnění jej vidět atp. V případě implementace v NDK bude DIP balíček pro uživatele dostupný pouze na vyžádání přes administrátora a bude obsahovat archivní digitální objekty a metadata. Běžné uživatelské kopie budou zpřístupňovány ze serverů mimo LTP systém a skrz aplikace zpřístupnění jako jsou Kramerius, WebArchiv a Manuscriptorium.

6.4 Koncept důvěryhodného digitálního repozitáře a jeho certifikace

S rozvojem využívání digitálních repozitářů se objevila otázka, jak prokázat, že konkrétní repozitář dodržuje určitá pravidla a procesy, které zaručují, že o data je a bude dobře postaráno. Jak si uživatel může být jistý, že dokument, který si prohlíží z digitálního repozitáře nebyl změněn, jak si může být jistý, že digitální repozitář nebude za měsíc zrušen kvůli nedostatku financí. Již v průkopnické zprávě *Preserving Digital Information* [GARRETT a WATERS, 1996] byla jako kritický komponent rozvoje digitálních archivů jasně popsána nutnost vytvoření dostatečně velké sítě certifikovaných archivů, které by zajistily odpovídající uložení digitálních dat, jejich migrace na nový HW i zpřístupnění. Bez certifikace nebudou producenti ani uživatelé repozitářům důvěřovat, že s daty pracují odpovídajícím způsobem a jsou schopny data dlouhodobě ochránit. To by znamenalo nejistotu ve financování, nejistotu migrací dat na nový HW apod. [GARRETT a WATERS, 1996, s. 40]. Zpráva zároveň rozebírala jednotlivé oblasti, které musí archiv (repozitář) řešit, aby mohl bezpečně uchovávat digitální objekty pro budoucnost. Uvedeny jsou mj. finance, strategie ochrany, výběr dokumentů, uvedeny jsou také modely finančních nákladů, které již tehdy byly známé (např. *Yale cost model*).
V roce 2002 přišlo RLG a OCLC s finální verzí zprávy nazvané *Trusted Digital Repositories: Attributes and Responsibilities* [RESEARCH LIBRARIES GROUP, 2002], která se jako první zabývala charakteristikami a odpovědnostmi, které se týkají digitálních repozitářů v kulturních institucích. Zpráva vznikla právě proto, že od poloviny 90. let 20. století vznikalo stále více a více digitálních repozitářů v institucích, které se tak stávaly zodpovědnými za dlouhodobý přístup ke světovému kulturnímu, sociálnímu a vědomostnímu dědictví. Důvěryhodný digitální repozitář tato zpráva definovala jako repozitář, „*jehož cíl je poskytovat spolehlivé, dlouhodobé zpřístupnění spravovaných digitálních zdrojů cílové komunitě, nyní i v budoucnu.*“ [RESEARCH LIBRARIES GROUP, 2002, s. i] V Německu v rámci programu *Nestor*[258] vznikla tato definice repozitáře: „*Repozitář je organizace (sestávající z lidí a technických systémů), která přijala odpovědnost za dlouhodobou ochranu digitálních objektů a za dlouhodobý přístup k nim, zajišťujíc jejich použitelnost určitou cílovou skupinou uživatelů.*“ [NESTOR, 2008, s. 4] Tato definice reflektuje, že repozitář není pouze HW, ale i prostředí okolo HW. Definice je tak velmi podobná definici OAIS archivu, která je součástí referenčního rámce OAIS – viz kapitola 3.3.1.

Důvěryhodnost je klíčovou vlastností, kterou musí certifikovaný repozitář demonstrovat. Certifikací se rozumí skutečnost, že podle konkrétních kritérií byla uznána (prokázána) schopnost repozitáře zachovat digitální objekty v dlouhodobém horizontu přístupné a použitelné. Takový repozitář se často nazývá důvěryhodný (v angličtině trustworthy). Důvěru lze obecně definovat jako: „*Spolehnutí na integritu, sílu, schopnost, záruku apod. u osoby nebo věci.*“ [Trust, 2011] V

[258] http://www.langzeitarchivierung.de/

případě repozitáře jde o důvěru uživatelů současných i budoucích, zřizovatele (financuje provoz), provozovatele, producentů dat i pracovníků v to, že systém funguje podle specifikace, jeho účelu a cílů a poskytuje důvěryhodný obsah. Jinými slovy systém dělá přesně to, co má dělat a na co byl vytvořen. Součástí důvěryhodnosti jsou bezpečnost, autenticita, integrita a dostupnost dat. Integrita znamená kompletnost a vyloučení jakýchkoliv nechtěných změn digitálního objektu, ať způsobených technickou chybou nebo záměrně člověkem. Autenticita zaručuje, že digitální objekt obsahuje to, co obsahovat má. Tato záruka je dána poskytnutím údajů o provenienci, o všech změnách objektu. Dostupnost dat spočívá v garanci archivu potencionálním uživatelům, že digitální objekty z archivu budou stále dostupné a použitelné. Důvěryhodnost se neomezuje na technické řešení, ale ve větší míře i další okolnosti, jakými jsou např. financování repozitáře, podmínky v jakých je provozováno, jak schopný má instituce personál, jaké má zabezpečení apod. [HUTAŘ, FOJTŮ a PAVLÁSKOVÁ, 2008, s. 2].

Obecné hledisko říká: Pokud je repozitář schopen rozpoznat a priorizovat hrozby, které by ohrozily jeho aktivity a vypořádat se s nimi tak, že sníží možnost jejich výskytu, k tomu navíc dokáže identifikovat možné „nepředvídatelné“ události, které opět představují jistou hrozbu, pak je velmi pravděpodobně připraven obdržet status důvěryhodnosti a získat tak případně certifikaci.
Certifikovaný repozitář by měl:

- přijmout odpovědnost za dlouhodobou péči o svěřené digitální objekty a za jejich zpřístupnění současným i budoucím uživatelům;
- organizačně zajistit dlouhodobou životnost nejen pro vlastní repozitář, ale i pro svěřené digitální informace;
- prokázat finanční zajištění i trvale udržitelný rozvoj.

K prokázání důvěryhodnosti a případné certifikaci vede cesta přes audit. Ten může být interní, pomocí dostupných nástrojů (např. DRAMBORA – viz kapitola 6.4.3.1), nebo může jít o nezávislý externí audit a následnou certifikaci. Důvody proč podstoupit interní nebo externí audite mohou být kromě prokázání důvěryhodnosti a získání certifikace různé. Může jít o externí potvrzení správnosti již provozovaných a nastavených procesů; o hledání slabých míst; může jít o získání podkladu pro další rozvoj. Audit je tedy určen provozovatelům repozitáře, jeho uživatelům a je velmi nápomocen i ve vztahu ke zřizovatelům a poskytovatelům financí, kterým je potřeba prokázat, že nastavený směr je dobrý a za své peníze dostávají to, co je třeba.
Jediný široce akceptovaný standard pro prokázání důvěryhodnosti neexistuje. Možností jak prověřit repozitář je více. Dostupné jsou standardy Nestor a TRAC postavené na bázi *checklistu* (tedy kontrolního seznamu), nebo flexibilní online nástroje jako je DRAMBORA. Celou oblast se snaží uchopit standardizační iniciativy, jako je např. skupina „*The Birds of a Feather Group*“[259], která se snažila o to, aby se TRAC stal ISO normou.

Vzhledem k rozvoji systémů na logickou dlouhodobou ochranu digitálních dokumentů a možnostem certifikace digitálních úložišť, které dávají vydavatelům jistotu, že jejich dokumenty jsou v dobrých rukou, roste v mnoha zemích zájem samotných vydavatelů o uložení jejich digitálních dokumentů v repozitářích provozovaných státní nebo národní knihovnou, případně archivem. V mnoha zemích to mají přímo za povinnost vyplývající z legislativní úpravy institutu

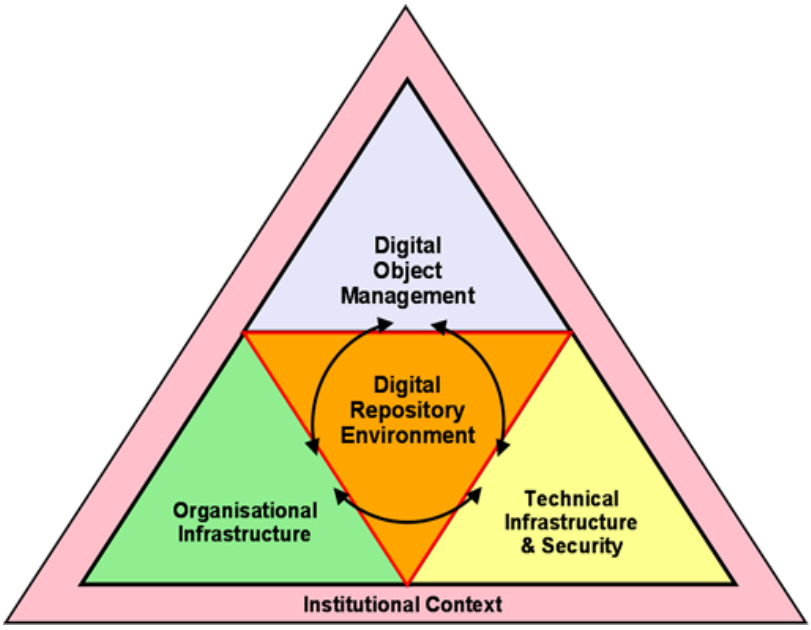
[259] http://wiki.digitalrepositoryauditandcertification.org/bin/view/Main/WebHome

povinného výtisku, který se vztahuje i na elektronické publikace. To, že by se dlouhodobá ochrana dat nechala na vydavatelích, je nemyslitelné. Naopak, ukazuje se, že právě zajištění ochrany jejich dat do budoucna v repozitáři knihovny může být lákadlem, na které vydavatelé slyší a jsou ochotni za něj poskytnout finance, případně omezený přístup ke svým dokumentům.

6.4.1 Deset základních principů důvěryhodnosti digitálního repozitáře

V lednu 2007 se v Chicagu v *Center for Research Libraries* sešli zástupci čtyř organizací věnujících se otázkám repozitářů a dlouhodobé ochraně digitálních dat, aby se pokusili nalézt deset klíčových kritérií hodnocení digitálních repozitářů. Jednalo se o zástupce organizace *The Digital Curation Center* (Velká Británie); *Center for Research Libraries* (USA); evropského projektu *DigitalPreservationEurope* a německé kooperační sítě pro dlouhodobou ochranu *Nestor*. Vzniklých deset kritérií se mělo stát základem pro další snahy na poli certifikace a auditů digitálních repozitářů. Těchto deset kritérií lze použít jako prvotní východisko vlastního ohodnocení předtím, než se opravdu přistoupí k serióznímu auditu nebo procesu certifikace. Můžeme je také považovat za základní popis obecných vlastností každého repozitáře, který chce být považován za důvěryhodný – viz Obrázek 20. Takový repozitář [CENTER FOR RESEARCH LIBRARIES, 2007]:

1. se musí zavázat k neustálému opatrování digitálních objektů pro určitou cílovou komunitu;
2. musí prokázat svou způsobilost a životaschopnost (včetně financování, personálních otázek, struktury, procesů), aby dostál stanovenému závazku;
3. musí si osvojit a dodržovat potřebná smluvní a legislativní pravidla a dostát všem z nich plynoucím závazkům;
4. musí mít efektivní a dostačující rámcovou strategii;
5. získává a ukládá digitální objekty na základě stanovených kritérií, které odpovídají cílům a schopnostem instituce;
6. neustále udržuje integritu, autenticitu a využitelnost digitálních objektů, které trvale uchovává;
7. vytváří a uchovává potřebná metadata o událostech souvisejících s uloženými digitálními objekty v průběhu jejich uchování, jakož i metadata o samotném vytvoření digitálních objektů, podmínkách zpřístupnění a kontextu využití digitálních objektů;
8. musí naplnit nezbytné požadavky na zpřístupnění objektů ven z repozitáře určité komunitě;
9. musí mít strategii pro plánování ochrany a souvisejících procesů včetně procesů pro krizové situace;
10. musí mít technickou infrastrukturu adekvátní pro účel neustálé údržby a zajištění digitálních objektů.



Obrázek 20 – Obecné vyjádření souvislostí 10 základních principů důvěryhodného repozitáře [ROSS, et al., 2009, s. 26].

6.4.2 Nástroje na externí audit digitálního repozitáře

Jak ale zjistíme, že konkrétní repozitář je důvěryhodný, nebo by se o toto označení mohl ucházet? Tady ke slovu přicházejí procesy auditu a certifikace.

6.4.2.1 TRAC (Trustworthy Repositories Audit & Certification)

Původ dnes nejznámější metodiky externí certifikace, známé pod zkratkou TRAC (*Trustworthy Repositories Audit & Certification*), lze hledat ve spojení RLG-NARA pracovní skupiny zabývající se výzkumem certifikací digitálních archivů s *Digital Curation Centre*[260] (DCC) z Velké Británie a s odborníky z německého projektu *Nestor*, které se odehrálo na počátku nového tisíciletí. Nejpodstatnější se ukázala spolupráce s *Center for Research Libraries* (CRL), která je dnes jedinou externí certifikační autoritou, která provádí audity a certifikace s metodikou TRAC. Právě CRL dostalo v roce 2005 finance z *Mellonovy nadace* na vývoj procesů a aktivit potřebných k auditu a certifikaci digitálních archivů. V rámci tohoto projektu CRL spolupracovala s RLG-NARA skupinou na předělání a doplnění již hotového dokumentu *Trusted Digital Repositories: Attributes and Responsibilities* [RESEARCH LIBRARIES GROUP, 2002]. Testování probíhalo ve spolupráci s britským DCC, které ve stejné době vyvíjelo nástroj pro interní audit DRAMBORA – viz kapitola 6.4.3.1. Výsledkem snah CRL a RLG byla v roce 2007 první verze dokumentu *Trustworthy Repositories Audit & Certification: Criteria and Checklist* [OCLC, CRL, 2007], který je dnes běžně označován jako „TRAC“ nebo „TRAC list“. Tento dokument poskytl metodiku pro audit, posouzení a certifikaci digitálního repozitáře. Zároveň vymezil i samotný postup auditu.
Starší z obou metodik certifikace dokumentů [RESEARCH LIBRARIES GROUP, 2002, s. 5] definuje obecné vlastnosti důvěryhodného repozitáře, z nichž některé se pak promítly do již zmíněných deseti principů důvěryhodného repozitáře – viz kapitola 6.4.1. Důvěryhodný repozitář tedy musí:

[260] http://www.dcc.ac.uk/

- přijmout odpovědnost za dlouhodobou péči o svěřené digitální objekty a za jejich zpřístupnění současným i budoucím uživatelům;
- organizačně zajistit dlouhodobou životnost nejen pro vlastní repozitář, ale i pro svěřené digitální informace;
- prokázat finanční zajištění v současnosti i trvale udržitelný rozvoj;
- navrhnout systém pro správu digitálního repozitáře v souladu s obecně platnými konvencemi a standardy v zájmu zaručení trvalé správy, zpřístupnění a zabezpečení uložených digitálních dokumentů;
- stanovit metodiku hodnocení důvěryhodnosti systému;
- jasně a srozumitelně prezentovat svoji odpovědnost za dlouhodobou ochranu a zpřístupnění dokumentů uživatelům i subjektům, které své dokumenty v úložišti deponují;
- disponovat strategií, pracovními postupy a službami, které umožňují snadné hodnocení a měření.

Navazující dokument *Trustworthy Repositories Audit & Certification: Criteria and Checklist* již představuje velmi detailní a propracovaný systém kriterií pro hodnocení důvěryhodného repozitáře. TRAC vychází z referenčního rámce OAIS, ze kterého přebral technikou část. OAIS považuje za vztažný bod pro úspěšné porovnávání. TRAC od počátku aspiroval na to stát se standardem a de-facto se jím stal. V současnosti jde o jediný dokument, podle kterého v oblasti digitálních knihoven a repozitářů opravdu probíhá externí certifikace, zakončená udělením certifikátu, zprávou o celém procesu apod. Celý proces certifikace probíhá na objednávku, třetí nezávislou stranou (auditorem). Tímto auditorem je v současnosti konsorcium CRL – *Centre for Research Libraries*[261]. Celý proces je samozřejmě placený, za jeden audit je uváděna cena mezi 50-70 tisíci USD. V poslední době získal TRAC certifikaci digitální repozitář organizace *HATHI Trust*. Samotný proces certifikace trvá většinou krátce přes jeden rok [HATHI TRUST, 2011], v případě repozitáře organizace *Portico*[262] v roce 2010 trval deset měsíců. Na podzim 2011 získal certifikaci TRAC repozitář Národní knihovny Nového Zélandu [KNIGHT, 2011]. Celý proces certifikace TRAC, jeho trvání, východiska, předpoklady a přínosy jsou detailně popsány v článku *Becoming a certified trustworthy digital repository: the Portico experience* [KIRCHHOFF, et al., 2010]. Pro všechny organizace platí, že než se pustí do externí certifikace, měly by si udělat tzv. *self-audit* (interní audit vlastními silami), který jim pomůže se připravit, shromáždit potřebnou dokumentaci všech procesů, ujasnit si procesy apod. Vlastní certifikace je ve formě stanoviska, zda konkrétní repozitář, jeho procesy a služby, i organizace okolo odpovídá (a do jaké míry), nárokům popsaným v metodice TRAC.

Hodnocení prováděné pomocí TRAC se zaměřuje na tři hlavní oblasti, které jsou dále rozpracované na okruhy a ty na dílčí kritéria:

1. **A – Organizace (řízení, struktura, udržitelnost, finance);**
   A1. Řízení organizace a její stabilita
   A2. Struktura organizace a zaměstnanci
   A3. Procedurální odpovědnost a strategický rámec
   A4. Finanční udržitelnost

[261] http://www.crl.edu/
[262] http://www.portico.org/digital-preservation/news-events/news/general-news/portico-certified-as-trustworthy-digital-repository-by-the-center-for-research-libraries

A5. Smlouvy, licence a závazky

2. **B – Správa digitálních objektů;**
   B1: Příjem dat: akvizice obsahu/dat
   B2: Příjem dat: vytvoření archivního balíčku
   B3: Plánování ochrany
   B4: Archivní uložení a ochrana /správa AIP balíčků
   B5: Information management
   B6: Správa přístupů
3. **C – Technologie, technická infrastruktura, bezpečnost**
   C1: Obecné nároky na infrastrukturu systému
   C2: Vhodné technologie, postavené na systémové infrastruktuře, další kritéria určující užití technologií a strategií vhodných pro tzv. cílové komunity
   C3: Bezpečnost IT systémů (servery, firewally, až po připravenost přírodní katastrofy)

V jednotlivých oblastech jsou sledovány okruhy, v jejichž rámci je třeba zodpovědět poměrně konkrétní otázky. Výsledné hodnocení je pro tyto okruhy v podobě bodů (tzv. úrovní certifikace). Body mohou být v rozmezí 1-5, kde 5 je nejvyšší úroveň a 1 je úroveň minimální, kterou lze certifikovat. Součástí výsledku je také vysvětlující text a zpráva o auditu. Vlastní audit neprobíhá pouze poměřováním procesů oproti nějaké sadě kritérií. Jde o to, zda repozitář opravdu dělá to, co říká, že dělá, zda k procesům existuje dokumentace a zda jde opravdu za cílem, který má stanoven. Každý repozitář je jiný, s různou úrovní dokumentace, procesy i cíly. I tak lze různé repozitáře certifikovat např. na stejné úrovni.
Pro celý proces je klíčová dokumentace všech procesů, politik, strategie a jejich dodržování. Právě dokumentace je prvním bodem, který slouží jako východisko pro CRL k auditu konkrétní organizace, jak uvádí i metodika TRAC [OCLC, CRL, 2007, s. 6]. V případě certifikace repozitáře *Portico*, bylo v úvodu auditu poskytnuto auditorům 1225 stran dokumentace [KIRCHHOFF, et al., 2010, s. 89], rozdělených do pěti skupin (organizace; strategie; systémová architektura a datový model; údržba a procesní a systémový vývoj; rozhranní). Vše popsané v dokumentaci je během auditu důkladně prověřeno na místě, na příkladech, případně přímo při provádění procesů.
Z uvedeného výčtu je zřejmé, že pro vybudování a provoz důvěryhodného repozitáře zdaleka nestačí zakoupení drahého technického a programového vybavení, právě naopak. Jedná se o složitý a dlouhodobý proces, který musí být řádně zakotven ve strategických prioritách i organizační struktuře instituce, která aspiruje na vybudování a provoz důvěryhodného digitálního repozitáře. Ten musí být adekvátně finančně a personálně zajištěn nejen pro dobu vzniku repozitáře a po krátkou dobu po něm, ale dlouhodobě. Tvorba a provoz důvěryhodných digitálních úložišť jsou po všech stránkách natolik naléhavé a zároveň náročné, že přesahují možnosti i těch největších institucí. Proto se jedná o velkou výzvu ke spolupráci paměťových a vědeckovýzkumných institucí v národním i mezinárodním kontextu [STOKLASOVÁ a HUTAŘ, 2007, s. 90].

Metodiku TRAC a další nástroje využila skupina odborníků, která si říká *International Audit and Certification Birds of a Feather Group*[263]. Jejich snahou bylo vytvoření ISO standardu, podle kterého by se v budoucnu prováděl kompletní audit a certifikace repozitáře. Existující nástroje se

[263] www.digitalrepositoryauditandcertification.org

proto snažili racionalizovat do jednoho dokumentu. Jejich snaha byla korunována úspěchem 14. 2. 2012, kdy návrh normy, tzv. purpurová kniha (*magenta book*)[264], byl po několikaletém procesu publikován jako finální norma ISO 16363[265]. Norma a její původní návrh vycházejí ve všech významných aspektech z metodiky TRAC, včetně struktury a většiny kritérií. Dokument poskytuje návod a metriku auditu, lze předpokládat, že bude využívána především na externí audity.

6.4.2.2 Data Seal of Approval (DSA)

Další možností získání certifikace je holandsko-německá iniciativa *Data Seal of Approval*[266] (DSA), která vznikla v roce 2009. Jde o metodiku sestávající ze 16 kritérií[267], které má repozitář ucházející se o certifikaci splnit. Certifikace se orientuje na repozitáře s výzkumnými daty, ale je použitelná na jakýkoliv repozitář. V 16 kritériích lze rozpoznat zmíněných 10 kritérií pro důvěryhodný repozitář – viz kapitola 6.4.1. Těchto 10 kritérií je dále doplněno. Celý proces probíhá tak, že instituce napřed provede dle kritérií vlastní interní audit. Pokud je s výsledkem spokojena, přihlásí se na certifikaci a DSA komise repozitář sama posoudí. Pokud je vše v pořádku, obdrží instituce provozující repozitář „pečeť kvality DSA“. Do dnešního dne certifikát získaly např. Národní knihovna Německa, UK Data Archiv a další. Pokud je repozitář a instituce připravena, jde do jisté míry o dostupný certifikát, který se stává výchozím bodem dalších certifikačních snah v EU.

6.4.3 Nástroje na interní audit digitálního repozitáře

Nástroje na interní audit jsou rozšířenější než metodiky na externí audity a certifikace. Důvodem je jejich podstata a určení pro potřebu institucí, které s jejich pomocí mohou svůj repozitář ohodnotit samy. Vlastní audit (*self-audit*) pomůže instituci poznat stav repozitáře, problémy a rizika a je tak výbornou přípravou na případný externí audit. Ten pak je o poznání kratší a jednodušší. K *self-auditům* vzniklo několik nástrojů a metodik, většina z nich v Evropě.

6.4.3.1 DRAMBORA

Nástroj DRAMBORA[268] (*Digital Repository Audit Method Based on Risk Assessment*) vznikl ve spolupráci *Digital Curation Centre* (DCC) a projektu *DigitalPreservationEurope* (DPE), kterého se účastnila i NK ČR. Díky tomu měla možnost se na vývoji a testování nástroje DRAMBORA v obou verzích účastnit. DRAMBORA nevznikla jako další certifikační nástroj, ale jako nástroj, který by měl pomoci instituci, která certifikaci svého repozitáře plánuje. Cíl nástroje DRAMBORA není hodnotit výsledky ostatních repozitářů, ale poskytnout nástroj k ohodnocení vlastního repozitáře a organizace. Pomoci může tím, že si instituce provede tzv. interní audit sama a odhalí slabiny a nedostatky organizační i týkající se repozitáře. Může to provést pracovník té instituce, musí ovšem být s nástrojem seznámen. Tento „samoodhad“ může do velké míry snížit náklady na následnou placenou externí certifikaci a také, což je podstatné, ji může významně urychlit. Je pouze na instituci, zda zvolí variantu, kdy audit bude svěřen internímu zaměstnanci nebo externí osobě.

[264] http://public.ccsds.org/publications/archive/652x0m1.pdf
[265] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=56510
[266] http://datasealofapproval.org/
[267] http://datasealofapproval.org/?q=node/35
[268] http://www.repositoryaudit.eu

První možnost je výhodná díky tomu, že zaměstnanec zná poměry a procesy v repozitáři, na druhou stranu může mnohá fakta záměrně či nevědomě opomíjet. Externí osoba bude objektivnější, ale audit bude trvat pravděpodobně déle.

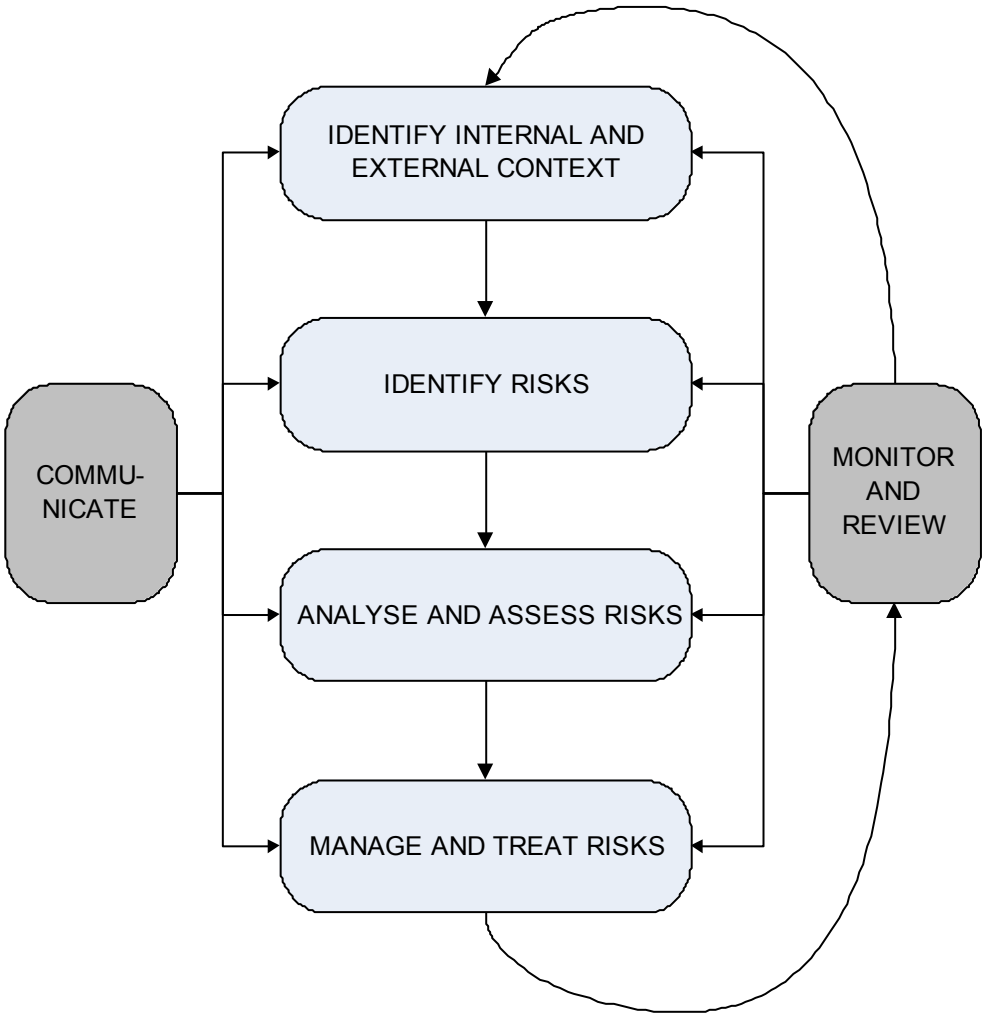
DRAMBORA je strukturována jako možná odpověď nebo pomůcka pro další vývoj úložišť a hlavně pro jejich funkčnost v nedaleké i daleké budoucnosti. DRAMBORU lze použít i jako nástroj plánování při budování repozitáře, kdy je možné se vyvarovat nesrovnalostí a ověřit, zda plán má vše potřebné. I proto má DRAMBORA záměrně hodně styčných bodů s certifikačními příručkami TRAC i Nestor, aby se oba přístupy (*self-audit* a externí audit) vzájemně doplňovaly. Výhodou pro samotnou instituci je, že může audit díky interaktivnímu rozhraní, kde lze uložit a editovat všechny provedené audity, pravidelně opakovat. To je velmi dobré v tom případě, že instituce si udělá první audit svého repozitáře, odhalí jeho slabiny, ohodnotí rizika, jejich možné následky a přijme opatření na zlepšení konkrétních věcí do určitého data (např. do jednoho roku). Po uplynutí této doby může nechat repozitář projít znovu celým procesem a výsledky obou auditů lze porovnat. Již známá rizika budou pravděpodobně potlačena, nebo snížen jejich případný dopad, objeví se ovšem rizika nová.[269] Důvodem proč se instituce může rozhodnout pro interní audit, může být také ověření, zda všechny procesy a celková strategie jsou nastaveny dobře a systém pracuje optimálně.

Nástroj DRAMBORA pracuje s riziky, skutečnými i možnými, které repozitáři hrozí. Ochrana digitálních dat je v podstatě uvědomování si organizačních, procedurálních, technologických a jiných nejistot a jejich přeměna na konkrétní měřitelná a řešitelná rizika. Audit mj. pomůže:

- rozpoznat a stanovit prioritní (největší) rizika, která ohrožují aktivity repozitáře,
- vypořádat se s riziky tak, aby se snížila možnost jejich výskytu,
- identifikovat možné nepředvídatelné události, aby se snížil efekt rizik, které představují.

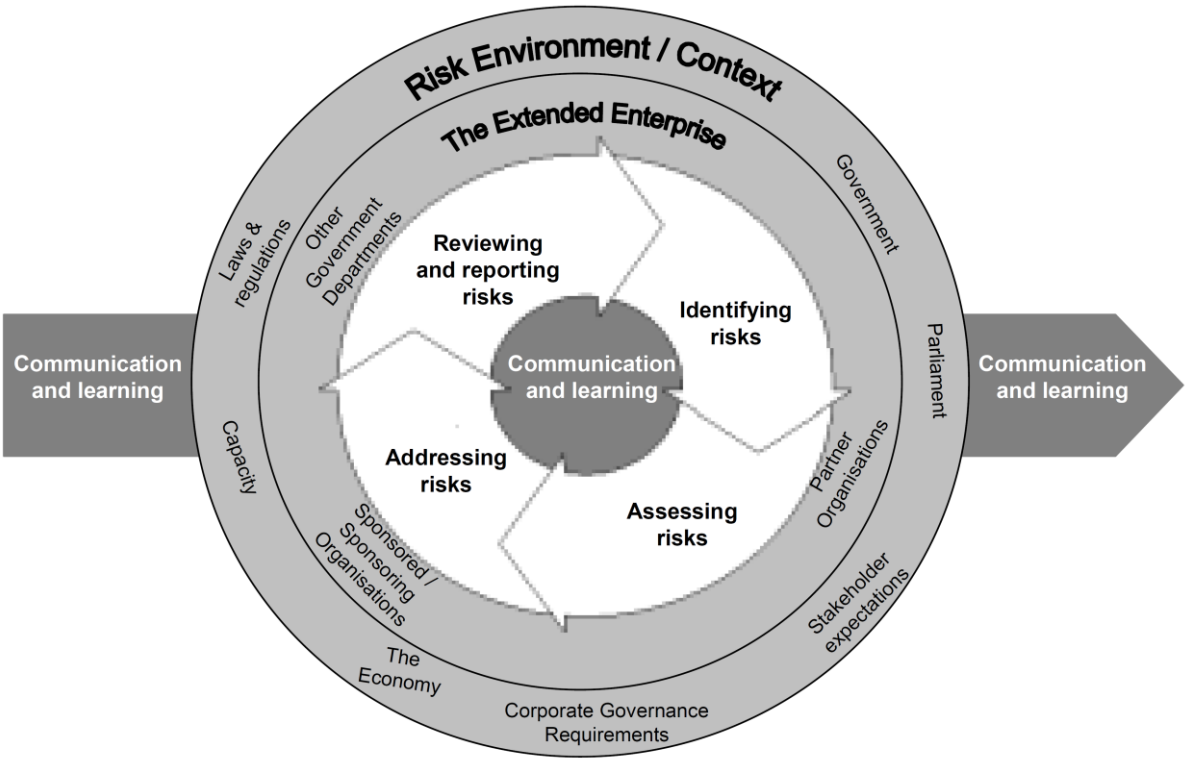
Poslední dva body platí zvláště pro nástroj DRAMBORA, který na rozdíl od ostatních metod a nástrojů auditu a certifikace (TRAC, Nestor) pomáhá s plánováním opatření na potlačení zjištěných rizik. Možná rizika, jejich dopad a pravděpodobnost výskytu jsou ohodnoceny číselnými hodnotami. Celkové bodové hodnocení rizik vychází z těchto hodnot a rizika jsou tak velmi dobře měřitelná a hlavně porovnatelná v čase. Grafické vyjádření opakovatelnosti procesů v nástroji DRAMBORA ukazuje Obrázek 21.

[269] V ČR tímto procesem s nástrojem DRAMBORA prochází jednou ročně Národní technická knihovna se svým repozitářem NUŠL – Národní úložiště šedé literatury.



Obrázek 21 – Grafické vyjádření procesů v nástroji na *self-audit* DRAMBORA [ROSS, et al., 2009, s. 13].

Dlouhodobá ochrana digitálních dat je *risk management* na všech stupních procesu rozloženého v čase [ROSS, et al., 2009, s. 5]. *Risk management* je proces měření nebo hodnocení risků a vytváření postupů jak je zvládnout. Pokud porovnáme procesy v auditu DRAMBORA s níže uvedeným Obrázek 22 znázorňujícím obecné procesy *risk managementu*, vidíme, že oba postupy jsou víceméně totožné. Obsahují koloběh začínající identifikací rizik, pokračující jejich řešením a končící hodnocením přijatých opatření.



Obrázek 22 – Management of Risk – Principles and Concepts [HM TREASURY, 2004, s. 13].

DRAMBORA v době svého vzniku v roce 2007, byla označena jako verze 1 a nebyla online nástrojem, jakým je dnes. Sestávala z metodického dokumentu a tabulek k ručnímu vyplnění. To práci na auditu velmi ztěžovalo. V auditu provedeném v NK ČR v létě 2007 narostl výsledný dokument ve formátu MS Word až na 100 stran. S verzí 1 bylo provedeno v roce 2007 celkem 8 pilotních auditů v institucích partnerů projektu *DigitalPreservationEurope*, všechny návrhy na zlepšení z nich vzešlé byly implementovány do DRAMBORA verze 2 dostupné od dubna 2008, která byla již online interaktivním nástrojem. Ten umožňuje audity ukládat, vracet se k nim, sdílet. Zároveň obsahuje databázi již hotových auditů a nápovědu, tvořenou právě vyplněnými hodnotami v různých úkolech předchozích auditů. Ve verzi 2 došlo nejen k přepracování celkového přístupu k auditu, ale i k několika úpravám v samotné metodice. Změnilo se číselné hodnocení pravděpodobností výskytu rizika a jeho dopadu, proto se může stát, že výsledky z DRAMBORY verze 2 jsou mírnější než z verze 1. Nejzásadnější změnou je zvýšení počtu tzv. funkčních tříd z 8 na 10. Ovšem nebylo to pouze přidání dvou tříd navíc, ale jejich celkové přepracování, tj. sloučení a naopak vyčlenění původních tříd [HUTAŘ, FOJTŮ a PAVLÁSKOVÁ, 2008]. Funkční třídy, které se dělí na procesní nebo podpůrné, jsou vlastně tematickými okruhy, vůči kterým se vypracovávají jednotlivé úkoly.

Procesní funkční třídy

- Akvizice a vstup digitálních objektů do repozitáře (*Ingest*)
- Ochrana integrity, autenticity a použitelnosti digitálních objektů
- Management metadat a kontrola průběhu zpracování
- Zpřístupnění dokumentů

- Plánování dlouhodobé ochrany digitálních dat a její aktivity

Podpůrné funkční třídy

- Mandát a závazek odpovědnosti ke správě digitálních objektů
- Stabilita a stav organizace
- Právní a regulační legitimita
- Účinné a efektivní strategie
- Odpovídající technická infrastruktura

Je možné využít všech deset tříd, nebo pouze ty, které jsou vypovídající pro konkrétní instituci nebo cíl auditu.

DRAMBORA audit má 6 fází, ve kterých je rozloženo 10 úkolů. Deset úkolů v rámci zmíněných šesti fází není rozmístěno rovnoměrně. Nejvíce úkolů se objevuje v první fázi (dva úkoly) a v druhé fázi (čtyři úkoly), v ostatních fázích je pak pouze jeden úkol. Níže je uveden přehled jednotlivých fází a úkolů.

Fáze 1 – identifikace východisek (kontext organizace/instituce)

- **Úkol 1: Specifikujte mandát vašeho repozitáře nebo organizace, kde je repozitář provozován.**
- **Úkol 2: Vyjmenujte cíle a účel repozitáře/organizace.**

  Tento úkol lze vztáhnout buď na všech 10 funkčních tříd, nebo pouze na ty, které jsou relevantní pro konkrétní audit. Ve výsledku tedy dostáváme rizika, která jsou rozdělena podle těchto tříd. Jsou tak daleko přehlednější a dá se na ně lépe reagovat konkrétními nápravnými opatřeními.

Fáze 2 – identifikace strategií a regulačního rámce

- **Úkol 3: Vyjmenujte dokumenty, které se týkají strategického plánování vašeho repozitáře.**
- **Úkol 4: Uveďte všechny právní, smluvní a regulační rámce nebo dohody, které se byť i jen okrajově dotýkají vašeho repozitáře.**

  Může jít např. o smlouvy s třetí stranou, zákony (např. zákoník práce, knihovní zákon, legislativní nařízení, autorský zákon, zákon o povinném výtisku, normy apod.).
- **Úkol 5: Vyjmenujte dobrovolná nařízení, standardy a manuály, ke kterým se repozitář hlásí a řídí se jimi.**

  Může jít např. o nekodifikované postupy např. vnitřní nařízení, předpisy atd.
- **Úkol 6: Vyjmenujte ostatní dokumenty a principy, se kterými je repozitář „ve shodě“.**

  Jakékoliv jiné dokumenty a skutečnosti, které ovlivňují chod a provoz repozitáře.

Fáze 3 – identifikace aktivit a prostředků

- **Úkol 7: Vyjmenujte všechny aktivity, prostředky potřebné k jejich provedení a jejich „majitele“.**

  Majitelem se rozumí osoba odpovědná za aktivitu, předmět této aktivity apod. Aktivitou se rozumí procesy probíhající v repozitáři tak, aby plnil svůj cíl a účel v rámci vnějšího kontextu. Tento úkol je již konkrétní analýzou aktivit a pracovních procesů.

Fáze 4 – identifikace rizik souvisejících s aktivitami a prostředky

- **Úkol 8: Identifikujte rizika spojená s aktivitami a prostředky vašeho repozitáře.**

  V tomto úkolu jde o popsání zranitelnosti a rizik, které se pojí ke všem aktivitám a procesům uvedených v předešlém úkolu. Je důležité si uvědomit, že jedno riziko ovlivňuje druhé a repozitáři hrozí vnitřní a vnější rizika. Je nutné škálovat rizika podle jejich možného „šířivého" efektu na své okolí. Tento efekt lze ohodnotit jako výbušný (*explosive*), nakažlivý (*contagious*), sdružený (*complementary*) a protikladný (*inverted*).[270]

Fáze 5 – vyhodnocení rizik

- **Úkol 9: Vyhodnoťte identifikovaná rizika.**

  Rizika se hodnotí číselnými hodnotami v určitém rozmezí. Je nutné vybrat, pro jakou z deseti uvedených tříd se riziko hodnotí. Hodnotí se pravděpodobnost výskytu (realizace) rizika a jeho případný dopad na třídu, kde by výskyt rizika udeřil nejvíce. Takovéto hodnocení se musí dělat pro každé riziko, které bylo během auditu identifikováno.

  Hodnocení pravděpodobnosti výskytu rizika se vyjadřuje číselnou stupnicí od 1 do 5.
  1 = *very low* – minimální pravděpodobnost, hrozba se objeví jednou za 10 let nebo méně.
  2 = *low* – pravděpodobnost nízká, jednou za 5 let.
  3 = *medium* – pravděpodobnost střední, jednou za 1 rok.
  4 = *high* – pravděpodobnost vysoká, výskyt jednou za měsíc.
  5 = *very high* – pravděpodobnost velmi vysoká, vyskytuje se vícekrát měsíčně.

  Hodnocení dopadu při výskytu rizika se vyjadřuje taktéž stupnicí od 1 do 5.
  1 = *very low* – zanedbatelný dopad; ústí v izolované, méně vážné a obnovitelné poškození nebo ztráty.
  2 = *low* – povrchní dopad; ústí v izolované, méně vážné a obnovitelné poškození/ztráty.
  3 = *medium* – střední dopad; ústí v poškození rozsáhlé nebo jdoucí napříč institucí; ztráty a/nebo poškození nejsou vážné a lze je vrátit do původní podoby, případně nahradit poškozená/ztracená data.
  4 = *high* – značný dopad; ústí v široké a vážné poškození nebo ztráty, které jsou nevratné nebo vratné pouze za pomoci třetí strany.
  5= *very high* – katastrofický dopad; ústí v nevratné a neopravitelné poškození a ztráty napříč institucí.

  Výsledné bodové hodnocení je součin pravděpodobnosti, že hypotetické riziko se stane realitou, a jeho očekávaného dopadu.

Fáze 6 – zvládnutí rizik

- **Úkol 10: Provádění opatření pro jednotlivá objevená rizika.**

  Cílem je vypracovat pro každé identifikované riziko postup, jak se vyhnout okolnostem, při kterých se rizika mohou objevit a tak snížit pravděpodobnost jejich výskytu a také snížit jejich potencionální dopad. Možnosti opatření jsou: vyhnutí se riziku (*avoidance*), tedy

[270] U verze 2 nástroje DRAMBORA došlo k vymazání nulového rizika, takové riziko nemá smysl vyjadřovat.

nastavení procesu vedoucího k tomu, aby se riziko nevyskytlo; a potlačení rizika (*treatment*), které spočívá v nastavení procesů k potlačení rizika, které se již vyskytuje.

Výstupem interního auditu DRAMBORA je celkový přehled procesů probíhajících v repozitáři a kolem něj. Přehled obsahuje seznam rizik včetně bodového ohodnocení těchto rizik. Na první pohled jsou tak patrné největší slabiny i silná místa repozitáře nebo jeho organizačního zajištění. To umožní urychleně a hlavně cíleně reagovat a možné problémy napravit dříve, než se stanou problémy či hrozbami skutečnými. Zároveň takto pojatý přehled procesů umožní pochopit více do hloubky samotné fungování repozitáře [HUTAŘ, FOJTŮ a PAVLÁSKOVÁ, 2008]. Záleží na auditorovi, jak podrobný výsledný přehled bude a také na instituci, jak kvalitní podklady a podporu auditorovi poskytne, což může ovlivnit jeho pochopení celého chodu repozitáře, potažmo instituce.
A konečně poslední výstup, kvůli kterému bude velká většina institucí *self-audit* podstupovat, je připravenost na kompletní externí audit (certifikaci).

Výsledky interního auditu mohou být překvapivé. Velmi často se za nejrizikovější oblast považují technologie, ovšem daleko větší riziko může být ukryto v organizačních a podpůrných procesech, které mohou potenciálně ochromit chod celého systému stejným způsobem jako technická závada. Z výsledků DRAMBORA auditů provedených v NK ČR v roce 2007 [HUTAŘ, 2008c], na ÚVT UK v roce 2008 [HUTAŘ, FOJTŮ a PAVLÁSKOVÁ, 2008] a v NTK, kde od roku 2009 probíhá každoročně [NÁRODNÍ TECHNICKÁ KNIHOVNA, 2011], je jasně patrné, že zásadním problémem není zakoupit a rozběhnout technologie k provozování repozitáře, ani je udržovat v chodu. Nejvážnějším problémem je organizační stránka instituce, která je hlavním předpokladem k odpovídající funkčnosti a existenci celého repozitáře. Právě organizační, finanční a personální zázemí jsou nejdůležitější k tomu, aby repozitář byl schopen plnit svou funkci – tj. uchovávat digitální objekty v dlouhodobém horizontu a zajišťovat k nim odpovídající přístup. Dalším problémem se ukazuje nastavení procesů, jejich dodržování a dokumentace; dostupnost a existence nutných metadat k zachování autenticity a integrity archivních objektů – viz [HUTAŘ, 2008c].
Pozitivní na provedených auditech je jednoznačně to, že si pracovníci uvedených institucí uvědomili slabiny, které jejich repozitář má. Věděli o nich již předtím, audit jim je ale potvrdil a bodově ohodnotil a v přehledné formě, kde je každé riziko popsáno včetně možných řešení. Jde o dokument, se kterým lze velmi dobře pracovat na potlačení většiny rizik nebo alespoň na snížení pravděpodobnosti jejich výskytu. Již během auditu vyšlo najevo, že v NK ČR i na UVT UK chybí v podstatě jakékoliv písemné podchycení procesů, a to i těch klíčových, které v repozitářích probíhají.

6.4.3.2 Nestor Criteria Catalogue

Nestor katalog kritérií je německou reakcí na aktivity okolo certifikace repozitářů. Již v roce 2004 vznikla v síti odborných institucí zabývajících se dlouhodobou ochranou digitálních dat *Nestor*[271] (*Network of Expertise in Long-Term Storage of Digital Resources*) pracovní skupina na certifikaci důvěryhodných repozitářů. Hlavním důvodem potřeby dokumentu, který by stanovil kritéria kvality repozitářů byl obrovský boom, který zažívaly německé instituce s budováním repozitářů

[271] http://www.langzeitarchivierung.de/eng/

(univerzity, výzkumné instituce, vládní úřady). Odborníci z projektu *Nestor* chtěli poskytnout návodný dokument a přimět provozovatele repozitářů k určité kvalitě a používání aktuálních standardů. Bylo rozhodnuto, že dokument bude vycházet z dokumentu *Trusted Digital Repositories: Attributes and Responsibilities* [RESEARCH LIBRARIES GROUP, 2002], ze kterého později vznikl TRAC. Nestor katalog kritérií je tak vlastně TRAC přizpůsobený německému prostředí (právnímu, finančnímu i celé komunity). Terminologie celého dokumentu se v maximální míře inspiruje referenčním rámcem OAIS. Veřejný draft ke komentářům spatřil světlo světa v roce 2006, po revizích a úpravách vyšla v roce 2008 finální podoba v němčině, anglický překlad v roce 2009 [NESTOR, 2009].

Katalog kritérií dělí celou problematiku do tří skupin:

- organizační rámec,
- správa digitálních objektů,
- infrastruktura a bezpečnost.

Skupiny dohromady obsahují 14 hlavních kritérií, z nichž většina se ještě dále dělí na další konkrétnější kritéria. U každého kritéria název, popis, příklad a přehled literatury, která se konkrétnímu aspektu věnuje. Hlavní kritéria jsou velmi abstraktní, důvodem je široký záběr katalogu určeného pro různé instituce a taky záměr, aby hlavní kritéria zůstala použitelná i v budoucnu.

Přesto, že vychází katalog kritérií Nestor z TRAC, jedná se čistě o metodiku na interní audit repozitáře pro paměťové instituce a univerzity a také o návod na jeho budování pro vývojáře, IT odborníky. Nejde o nástroj pro certifikaci, neexistuje žádná certifikační autorita.

6.4.3.3 Certifikace repozitářů v EU – výhled

Protože se ukázalo, že v Evropě a v zámoří pracují skupiny, které mají stejný cíl, tedy specifikaci kritérií pro důvěryhodný repozitář a proces certifikace, vznikla iniciativa, která tyto aktivity spojuje. Dne 8. 7. 2010 bylo podepsáno tzv. MoU (*Memorandum of Understanding*), čili dohoda o porozumění a spolupráci, mezi iniciativou *Data Seal of Approval*; americkou pracovní skupinou *Repository Audit and Certification* z CCSDS a německou pracovní skupinou *Trustworthy Archives – Certification*. Cílem dohody bylo připravit prostředí pro spolupráci těchto skupin s cílem vytvoření integrovaného rámce pro audity a certifikaci digitálních repozitářů [Trusted Digital Repository, 2010]. Integrovaný rámec pro certifikaci sestává ze tří úrovní certifikace, ve kterých se důvěryhodnost repozitáře zvyšuje:

- ***základní certifikace*** – je přidělena repozitáři, který obdrží DSA certifikát – viz kapitola 6.4.2.2;
- ***rozšířená certifikace*** – je přidělena repozitáři, který má certifikaci základní a k ní navíc provede strukturovaný, externě dohlížený a veřejně dostupný interní audit založený na ISO 16363 nebo německé normě DIN 31644;
- ***formální certifikace*** – je přiznána repozitáři, který navíc k základní certifikaci projde plným externím auditem a certifikací na bázi ISO 16363 nebo DIN 31644.

K celému systému certifikací bude existovat grafický symbol, který pak repozitář může používat. Za celou aktivitou vytvoření rámce pro certifikace stáli lidé, kteří připravovali ISO normu 16363 a také EU, která v další vlně projektů v 8. rámcovém programu (FP8) počítá s tím, že data vzniklá v tomto rámcovém programu budou muset být uložena v certifikovaných repozitářích. Proto je tlak na vznik postupů a typů certifikací.

6.4.4 Plánování a budování repozitáře – metodika PLATTER

Při plánování repozitáře je nutné se zamyslet nad otázkami, které mohou pomoci celkovému výsledku. Především je nutné vědět, proč repozitář potřebujeme, a zda vůbec. Dále je dobré zjistit, jaké technologie jsou dostupné, s jakými systémy chceme, aby repozitář spolupracoval, jaké znalosti musí mít zaměstnanci, zda takové zaměstnance máme nebo jsme schopni je najít. Je vhodné zkusit naplánovat růst objemů dat a možnosti rozšíření repozitáře. Významnou kapitolou je financování, které musí být jasné na mnoho let dopředu. Pokud máme toto vše rozmyšlené, je to dobrý výchozí bod pro budování repozitáře a záruka, že nás v budoucnu nic nepřekvapí tak, aby instituce nedokázala zareagovat a ohrozila tím repozitář. Velmi dobrý seznam bodů k zamyšlení pro jednotlivé části repozitáře přináší [REESE, 2008, s. 56]. Do úvahy by měly být vzaty i existující relevantní standardy a normy, jako jsou např. norma ISO 9000 (*Quality assurance*); ISO 17799 & 27001 (*Information security*), ISO 15489 (*Institutional Records Management*), ISO 14721 (referenční model OAIS).

Pro plánování repozitáře, který má být důvěryhodný a chystá se výhledově projít certifikací, vznikla v roce 2008 v rámci projektu DPE speciální metodika se zkratkou PLATTER (*Planning Tool for Trusted Electronic Repositories*), která je od roku 2009 dostupná v českém překladu [ROSENTHAL, BLEKINGE-RASMUSSEN a HUTAŘ, 2009]. Na přípravě publikace se díky účasti v projektu DPE podstatně podílela i NK ČR, konkrétně autor této práce. Text vznikal v rámci pracovního úkolu D3.2 pod názvem *Repository Planning Checklist and Guidance*[272]. Úkolu se kromě NK ČR účastnily Národní knihovna Dánska, Národní archiv Nizozemí a Univerzita Glasgow. PLATTER metodika vychází logicky z nástroje DRAMBORA, který byl také produktem projektu DPE a inspiraci si bere v metodice TRAC i v německém katalogu kritérií Nestor. PLATTER je určen pro správce (kurátory) digitálních dokumentů hledající návod, jak postupovat při vytváření digitálního repozitáře, který by po jeho zprovoznění bylo možné považovat za „důvěryhodný“ a v případě potřeby provést certifikaci. Smyslem použití PLATTERu ovšem může být také prostá snaha mít repozitář, který je funkční z dlouhodobého hlediska a má podchycené všechny eventuality, které mohou nastat. Metodika umožňuje provést plánování bez větší znalosti problematiky dlouhodobé ochrany digitálních dat a také s minimálními náklady finančními.

Projekt DPE chtěl poskytnout metodiku plánování, která ukáže proces definice cílů a procesů repozitáře z hlediska „důvěry“. PLATTER tedy není dalším nástrojem na audit a certifikaci, je spíše jejich doplňkem. Jde o průvodce plánováním budování digitálního repozitáře, který se zaměřuje na proces definice jeho cílů, než na popis toho, jak cílů dosáhnout. Při vytváření PLATTER bylo nutno se vyrovnat s tím, že bude využíván pro různé typy repozitářů. V úvodní fázi plánování je proto nutno pomocí PLATTERu specifikovat, o jaký repozitář se bude jednat. Specifikace probíhá pomocí klasifikačního dotazníku, který posuzuje repozitář z pohledu jeho účel a funkce; velikosti; nároků na provoz; nároků technické řešení a implementaci [HUTAŘ a ROSENTHAL, 2008]. Další postup metodiky PLATTER je následující:

- definovat obecné klíčové principy (platné pro všechny typy repozitářů);
- použít tyto jako základ pro stanovení vlastních cílů konkrétního repozitáře;
- poskytnout dost příkladů k použití, modifikaci (závisí na specifickém mandátu repozitáře).

[272] http://www.digitalpreservationeurope.eu/platter.pdf

PLATTER staví celý proces plánování na deseti klíčových principech důvěryhodného repozitáře (viz kapitola 6.4.1), ze kterých vzniklo 9 tzv. *Plánů strategických cílů* jako výchozí bod plánování [ROSENTHAL, BLEKINGE-RASMUSSEN a HUTAŘ, 2009, s. 21]. Jde o:

- Finanční plán,
- Akviziční plán,
- Plán řízení lidských zdrojů,
- Plán zpřístupňování,
- Technický plán,
- Datový plán,
- Plán zajištění kontinuity,
- Krizový plán,
- Plán ochrany.

Každý z těchto strategických cílů obsahuje dílčí úkoly a klíčové indikátory, které při finálním splnění v hotovém repozitáři pomáhají k tomu, že repozitář odpovídá deseti kritériím důvěryhodného repozitáře. Je pak pravděpodobné, že bude za důvěryhodný považován a může se ucházet o certifikaci.

7. Aplikační metadatový profil pro digitalizaci v projektu NDK

Již v předchozích kapitolách bylo uvedeno, že většině digitálních objektů ukládaných dnes v repozitářích chybějí právě ochranná a administrativní metadata, která jsou pro logickou dlouhodobou ochranu digitálních dokumentů tak důležitá. Knihovnická komunita je zvyklá se zabývat pouze popisnými metadaty, ostatní typy metadat donedávna považovala za nedůležité. Přitom přidání dalších metadat v prvních krocích vytváření nebo vzniku digitálního objektu, tj. v prvních fázích jeho životního cyklu, velmi podstatně přispívá k možnostem logické dlouhodobé ochrany dat i správě digitálních objektů. Repozitář musí zaručit, že digitální objekty jsou autentickými verzemi nějakého konkrétně identifikovatelného digitálního objektu první instance, jehož původ je jasně dokumentovaný. Proto je důležité mít údaje o procesech, které se odehrály při vzniku objektu, i jeho technické vlastnosti, které lze při uložení do repozitáře ověřit a dále doplnit. Dle toho vznikal i profil metadat pro digitalizaci v projektu NDK.

Specifikace metadatového profilu pro projekt *Národní digitální knihovna* začala v koncepční rovině již v roce 2007. Tehdy se začaly implementovat standardy PREMIS, METS, MIX do stávajících procesů digitalizace. A již tehdy šlo o tak rozšířené standardy, že se počítalo s jejich využitím pro projekt NDK. Opravdová práce na specifikaci nové sady a implementace začala v roce 2009, kdy vznikl první strohý popis konceptu tvorby metadat v digitalizaci NDK pro studii proveditelnosti. Jakmile byl projekt NDK na jaře 2010 schválen, začalo se s rozpracováváním všech standardů, s debatami s lidmi odpovědnými za digitalizaci, s firmami. Bylo potřeba udělat analýzy dostupných podobných řešení nasazení zmíněných standardů. Za referenční byly zvoleny digitalizační procesy a tvorba metadat v národních knihovnách Finska, Norska, Nizozemí, Kongresové knihovny USA a australské národní knihovny. Ve všech případech jde o instituce, které používají všechny jmenované standardy a ukládají vzniklé digitální objekty a metadata do LTP systému s funkcionalitou podobnou té, která byla plánována pro projekt NDK. S odborníky z některých těchto knihoven byl metadatový profil konzultován.

Cílem profilu je vytvořit v digitalizaci dostatečnou množinu administrativních, ochranných, technických a popisných metadat tak, aby z toho maximální mírou těžil LTP systém a tato metadata dokladující první okamžiky existence digitálního objektu uložil a využil pro procesy logické dlouhodobé ochrany. Profil není specifikací metadat na uložení v LTP systému, ani specifikace metadat pro vyhledávání a zpřístupnění digitálních objektů.[273] Jde čistě o specifikaci metadat, která mají vzniknout v digitalizaci. LTP systémy mají vnitřní formát, který odráží konkrétní datový model.[274] Ten by měl být schopen pojmout jakákoliv metadata v jakékoliv struktuře a standardu, samozřejmě po namapování na vnitřní formát LTP systému. Úlohou tvorby metadat v digitalizaci dle profilu NDK je to, aby bylo co do vnitřního modelu metadat LTP systému plnit.

[273] I když i pro oba procesy lze specifikaci, resp. obsah elementů využít, ne ovšem strukturu balíčku apod.

[274] Důvodem je to, aby metadata v LTP byla stále konzistentní, aby vnitřní formát udržel údaje z jakýchkoliv metadat a bylo možné doplňovat další údaje (např. události, změny aj.) stále do stejného (tedy vnitřního) formátu. Není možné do LTP ukládat jednou Dublin Core, podruhé MODS, potřetí MARXML, znemožnilo by to LTP systému práci a procesy, na které je určen.

Kontejnerové schéma METS bylo zvoleno proto, že je vhodné pro ukládání celého balíčku dat i metadat konkrétní intelektuální entity i mimo LTP systém. Knihovny digitalizující ve VISK7 tak mají možnost uložit svá data a metadata ve standardní podobě i když nemají LTP systém. Jejich data budou připravena na budoucnost. Použití zvolených standardů, XML syntaxe, i struktura balíčku z digitalizace dávají záruku, že metadata budou nezávislá na HW a SW, interoperabilní, a použitelná v různých systémech i procesech.
Specifikace profilu metadat pro projekt NDK byla součástí textu zadávací dokumentace, jako závazná příloha č. 6 [NÁRODNÍ KNIHOVNA ČR, 2011] pro výběr digitalizačního workflow, které museli zájemci o provedení implementace vyhovět. Text byl dopracován v únoru 2011, tendr byl vyhlášen až v červenci 2011. Mezitím byla specifikace vylepšována, především v souvislosti s projektem ANL+. Tento projekt má za cíl digitalizovat a popisovat analytické materiály, jako jsou např. články. Bylo rozhodnuto, že metadata budou shodná s projektem NDK, kde se také výhledově se zpracováním článků počítá. ANL+ od podzimu 2011 probíhá a metadata vznikají dle poslední verze specifikace standardu NDK, která je i součástí této práce. Projekt ANL+ má trvat několik let.
V následujících kapitolách jsou popsány standardy, které profil metadat pro NDK obsahuje, jeho cíle, některé aspekty a procesy, které vedly k jeho finální podobě. Vlastní návrh profilů je v příloze.

7.1 Jak vzniká aplikační metadatový profil

Rozhodnutí o používaných standardech v konkrétním projektu není zdaleka jednoduché. Je potřeba se rozhodnout, jaké metadatové schéma nebo schémata budou nejlépe vyhovovat cílům projektu, kterým může být prostý popis, dlouhodobá ochrana nebo další věci. Pokud se vybere schéma nevhodné, neslouží dobře ani uživatelům, ani administrátorům obsahu repozitáře, ani digitálnímu objektu samotnému.
Metadatový profil lze vytvářet z elementů různých schémat nebo jednoho schématu jejich kombinací tak, aby byl výsledek optimální pro konkrétní využití (např. Dublin Core je často doplněn elementy z jiných schémat). Lze také vytvořit schéma pouze z jednoho výchozího schématu změnou pravidel použití konkrétních elementů, případně jejich sémantiky a také doporučených kontrolovaných slovníků (viz např. schéma DC-lib založené na Dublin Core určené pro využití v knihovnách [ZENG a QIN, 2008, s. 114]). Poslední a nedoporučovanou možností je vytvořit aplikační profil doplněním existujícího schématu o elementy vlastní, nebo vytvoření kompletně nového schématu metadat. Profil NDK využívá první uvedené možnosti.

Důležité je i rozhodnutí o tom, která vlastnost metadat je nejpodstatnější pro dosažení našich cílů nebo cílů projektu samotného. Tj. určení granularity, hloubky popisu, podrobnosti záznamů. Rozsáhlost záznamu se velmi často odvíjí od časové a finanční náročnosti. Proto je potřeba najít kompromis mezi rozsáhlostí záznamu a skutečnými potřebami tak, aby záznam byl co nejpodrobnější při rozumné úrovni náročnosti časových a finančních nákladů na jeho vytvoření. Samozřejmostí by mělo být využití posledních verzí metadatových standardů, kontrolovaných slovníků (případně jejich vytvoření). Neměla by vznikat vlastní schémata, která mají spoustu omezení. V dnešní době je nutné dbát i na to, aby metadata vznikala s ohledem na archivaci a dlouhodobou ochranu popisovaných digitálních objektů, což je jeden ze záměrů profilu pro NDK. Ten se také snaží vyhovět principům tvorby metadat, které uvádí [COLE, 2002]:

- dobrá metadata musí být vhodně zvolena pro typ materiálu, který popisují;

- dobrá metadata podporují interoperabilitu;
- dobrá metadata mají specifikovány kontrolované slovníky;
- dobrá metadata podporují logickou dlouhodobou ochranu digitálních objektů.

Při vytváření návrhu implementace konkrétní specifikace nebo profilu metadat je dobré si promyslet několik otázek. Text níže částečně čerpá z [DAPPERT, 2009]. Obecné otázky, na které bychom měli znát odpovědi před specifikací metadat pro digitalizaci:

- pro jaké typy digitálních objektů chceme metadata vytvářet;
- co chceme popisovat – intelektuální entitu, soubor nebo bitstream;
- co je základní intelektuální entita, kterou budeme popisovat (titul, článek, kapitola aj.);
- co se je úkolem digitálních objektů (archivace, zpřístupnění aj.);
- jaké jsou vztahy mezi jednotlivými digitálními objekty/dokumenty/entitami;
- jaká metadata chceme uchovávat k zajištění dlouhodobé ochrany digitálních dat.

Pokud máme hotový plán, může se stát, že nebudeme z různých důvodů schopni jej naplnit, případně metadata dle našich představ vytvořit a využít. Aby k tomu nedošlo, musíme předem zvážit také následující skutečnosti:

- jaká metadata dokáže udržovat náš repozitář (LTP systém) a v jakém rozsahu;
- budeme-li popisovat Agenty, Události a Práva;
- existují-li již nějaká hotová metadata dostupná z jiných zdrojů (katalogy); pokud ano, jaké metadatové schéma je možné z katalogu exportovat;
- jaká technická metadata jsme schopni z digitálních objektů extrahovat;
- zda umí SW pro workflow digitalizace, který využíváme, vytvářet specifikovaná metadata;
- jaká bude struktura popisu.

Pakliže se rozhodneme o standardech, které budeme využívat, a o syntaxi záznamu metadat, je dalším krokem vytvoření specifikace využití konkrétních elementů jednotlivých standardů. Specifikace obsahuje návod a vysvětlení k celku i jednotlivým elementům; pravidla výskytu elementů a zápisu jejich hodnot (kontrolované slovníky, využívání autorit, pravidla popisu). Tak vznikne aplikační metadatový profil, který je promítnutím konkrétních potřeb projektu, objektů, komunity, nebo instituce do tvorby metadat.

V profilu pro NDK není zcela řešena otázka zápisu hodnot elementů, zvláště v popisných metadatech. Počítá se s automatickým plněním. Ovšem pokud bude nějaká instituce vytvářet dle profilu záznamy popisných metadat ručně, musí si vytvořit pravidla popisu pro katalogizátory, obdobná těm, která existují pro DTD monografie a periodika v NK ČR. Touto cestou se rozhodla jít Knihovna Akademie věd ČR, kde na přelomu let 2011 a 2012 Linda Jansová vytvořila pravidla popisu dle profilu NDK.

7.2 Metadatové standardy použité ve specifikaci metadatového profilu pro NDK

7.2.1 METS

V minulých letech vzniklo mnoho metadatových standardů pro objekty uložené v digitálních knihovnách, jediné co scházelo, byl celkový rámec, ve kterém by tato schémata mohla být integrována. METS je takovým rámcem. Všechna metadata navrhovaného profilu NDK budou

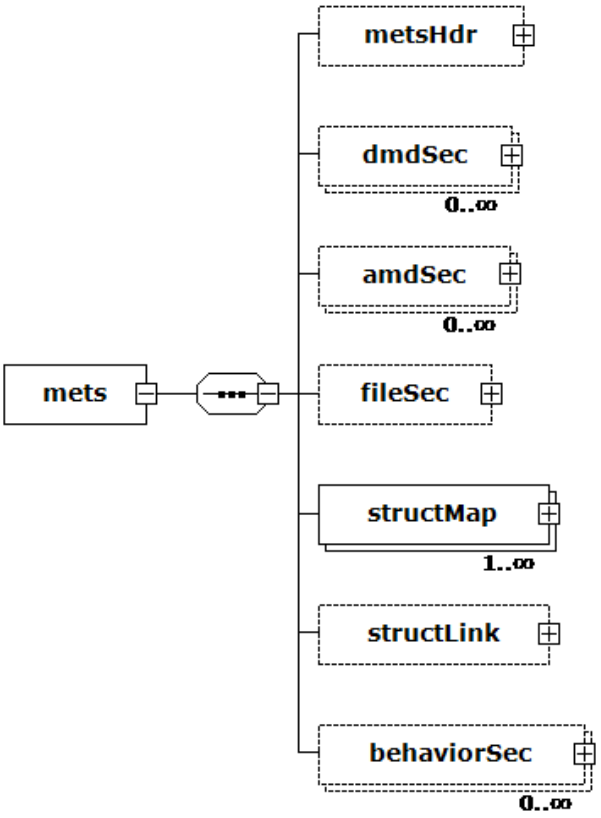
zabalena do kontejnerového standardu METS (*Metadata Encoding and Transmission Standard*), které je navržené pro zápis popisných, administrativních a strukturálních metadat udržovaných o objektech v rámci digitálních knihoven. METS vznikl za účelem vytváření XML dokumentů, které vyjadřují hierarchickou strukturu digitálního objektu, jména a lokace souborů, které tento objekt tvoří a s ním spojená metadata. Je tak tedy možné do METS záznamu vložit jeden nebo více dalších XML záznamů různých jiných schémat. Nástrojem pro vyjádření těchto metadat ve standardu METS je XML schéma (METS XSD[275]), které je v současnosti již delší dobu ve verzi 1.9. Vývoj schématu je popsán v kapitole 4.8.2. METS umožňuje vložit metadata přímo do své struktury (což je právě případ PREMISu), nebo do externích souborů, na které je z METS odkazováno. Profil NDK použije první variantu, tedy všechna metadata přímo v METS záznamu, bez odkazování mimo něj. Odkazování z METS záznamu na další metadatové záznamy se používalo např. v aplikaci Kramerius.

V NK ČR byl METS používán od roku 2007 do roku 2009 jako vnitřní formát systému Manuscriptorium. METS byl také implementován v aplikaci zpřístupnění Kramerius, aby bylo možno generovat METS záznamy jednotlivých úrovní konkrétních dokumentů (např. ročník konkrétního periodika) a použít jej jako výměnný formát – více viz kapitola 5.3.2.1. METS tedy nebyl nikdy v NK ČR použit jako kontejner na archivaci dat. Návrh profilu NDK s tímto použitím počítá, zejména v případech, že digitalizovaný dokument nebude uložen do LTP systému NK ČR, ale např. do repozitáře konkrétní knihovny. V tomto okamžiku bude uložen v METS podobě, která vznikla v digitalizaci a bude tak zajištěna interoperabilita a standardnost metadat i digitálního objektu do budoucna. V případě LTP systému NK ČR se počítá s tím, že METS bude fungovat jako transportní standard a pro uložení na dočasném pracovním prostoru předtím, než metadata, která obsahuje, budou namapována do vnitřního formátu LTP systému. Poté se METS vymaže. I tak ovšem zaručí, že během přenosu a dočasného uložení jsou data a metadata logicky spojena.

7.2.1.1 Standard METS a jeho části

METS jako kontejner pro vložení různých typů metadat sestává ze 7 hlavních částí. Celkový rámec zcela kompletního METS dokumentu vypadá tak, jak jej znázorňuje
Obrázek 23 níže. Povinná je pouze část <structMap> a kořenový element METS.

[275] http://www.loc.gov/standards/mets/mets.xsd



Obrázek 23 – Struktura záznamu METS.

Kde jednotlivé části jsou:

<METS:metsHdr> *hlavička*
<METS:dmdSec> *sekce popisných metadat*
<METS:amdSec> *sekce administrativních metadat*
<METS:fileSec> *seznam souborů dokumentu*
<METS:structMap> *sekce strukturálních map*
<METS:structLink> *sekce strukturálních linků*
<METS:behaviorSec> *sekce modelů chování*

Dvě z těchto METS sekcí nemají specifikován způsob zápisu. Jde o sekce popisných metadat <dmdSec> a o část administrativních metadat <amdSec>. Tyto sekce musejí být „naplněny“ již hotovým metadatovým popisem v jiném vhodném schématu nebo schématech. Návrh profilu pro NDK využívá všechny části, mimo dvou uvedených jako poslední (<StructLink> a <BehaviourSec>).

Níže uvedený popis jednotlivých částí METS standardu není vyčerpávající, pouze naznačuje možnosti schématu METS. Kompletní seznam atributů a pravidel je dostupný v METS XSD verze 1.9 nebo v dokumentu známém pod názvem *METS Primer* [DIGITAL LIBRARY FEDERATION, 2007].

kořenový element

V každém METS záznamu je nutný kořenový element <mets> s pěti možnými atributy (ID, OBJID, PROFILE, TYPE, LABEL). Poslední dva jmenované jsou využity v návrhu profilu NDK, který je součástí této práce. Popisují typ dokumentu, který je METS záznamem popsán a také jeho název nebo jiné označení. ID je vnitřní identifikátor části METS záznamu a vyskytuje se u většiny jeho částí. OBJID je identifikátor objektu, tj. konkrétního METS záznamu jako digitálního objektu.

Atribut PROFILE obsahuje informaci o tzv. METS profilu, pokud podle něj byl METS dokument vytvořen.
Aby se odlišily případné jednotlivé zdroje elementů se stejným jménem (např. author, compression aj.), jsou součástí kořenového elementu deklarace jmenných prostorů XML (xmlns). Každý jmenný prostor je identifikován pomocí URI. Musí být také připojen link na XML schématu použitém v METS záznamu, a to za pomoci atributu xsi:schemaLocation – viz uUkázka 7. Jde o výstup z masové digitalizace v NK Nizozemí v projektu Historické noviny[276].

```
<mets xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns="http://www.loc.gov/METS/" xsi:schemaLocation="http://www.loc.gov/METS/
http://schema.ccs-gmbh.com/METAe/mets-metae.xsd"
xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:mix="http://www.loc.gov/mix/" xmlns:xlink="http://www.w3.org/1999/xlink"
TYPE="Newspaper" LABEL="Haagsche courant no. 17460 06.01.1940">
```

Ukázka 7 – Kořenový element METS záznamu.

<metsHdr>

Hlavička popisuje vlastní METS dokument, jeho vznik, odpovědnou osobu apod. Podoba METS hlavičky vycházela ze schématu TEI, je ovšem o poznání stručnější. Může obsahovat atributy s daty vzniku nebo úpravy (CREATEDATE, LASTMODDATE) nebo atribut popisující stav záznamu (hotový, rozpracovaný, změněný apod. – RECORDSTATUS). V jedné z posledních verzí METS se objevila možnost vložit atribut ADMID, který linkuje na ID administrativní část metadat, která se týká samotného METS záznamu. Důležitý je dceřiný element <agent>, kam lze zapsat údaje o odpovědnosti za METS záznam (jméno a poznámku). Element <agent> má atribut ROLE, který má hotový slovník povolených hodnot (např. CREATOR, EDITOR, ARCHIVIST apod.).

<dmdSec>

Sekce popisných metadat k dokumentu, pro který byl METS záznam vytvořen. Může jít o popis předlohy i digitálního objektu (využíváno např. ve finské národní knihovně). Forma zápisu popisných metadat není ve schématu METS specifikována. Popisná metadata jsou definována jako část METS záznamu, kam se vkládá metadatový záznam jiného XML schématu (např. MODS). Využívá se stávajících schémat popisných metadat jako jsou DC, MODS, MARCXML apod. Případně lze nadefinovat vlastní. Každá sekce <dmdSec> musí mít své vlastní ID (jako atribut). Na toto ID pak odkazují jiné části METS záznamu (pomocí DMDID atributu). Je tak možné např. ke konkrétnímu souboru uvedenému v sekci <fileSec> udělat propojení na popisná metadata tohoto souboru (může jít např. o stránku knihy apod.). METS schéma, kvůli zaručení interoperability, doporučuje k využití následující schémata: MODS, EAD, VRA, Dublin Core, MARCXML, TEI část header, DDI (*Data Documentation Initiative*) a FGDC (*Federal Geographic Data Committee*). Vybrané schéma musí být uvedeno v atributu MDTYPE. Lze použít i jiná schémata, pak je v atributu hodnota OTHER. V Krameriovi se v první fázi používalo schéma MARCXML[277].

[276] http://kranten.kb.nl/

[277] Popisná metadata v tomto standardu se tvoří z konkrétních DTD jednotlivých dokumentů pomocí převodové tabulky.

Existují dvě možnosti, jak popisná metadata mohou být součástí konkrétního METS záznamu. Buď mohou být do METS záznamu vložená pomocí elementu <mdWrap>, nebo mohou být odkazovaná na záznam popisných metadat uložený jinde (mimo METS záznam). K tomu obecně slouží element <mdRef>. Lze také použít oba přístupy v jednom záznamu.

<amdSec>

Sekce administrativních metadat popisovaného objektu. Jednotlivé části (viz níže) popisují okolnosti vzniku objektu, jeho technické vlastnosti, podmínky využití, změny a události, které se digitálního objektu týkaly nebo týkají. Záznam o celém životním cyklu objektu je podstatný pro logickou ochranu digitálních dat a administrativní metadata jsou proto klíčová pro využití v LTP systémech. Forma administrativních metadat opět není specifikována a vkládají se XML záznamy jiných schémat. Nejčastěji se pro vyjádření administrativních metadat používá/integruje schéma PREMIS pro popis všech typů objektů a schéma MIX pro popis výhradně obrazových metadat. Element <amdSec> obsahuje dceřiné elementy, které tvoří čtyři podčásti sekce administrativních metadat. Jde o tyto elementy:

- <techMD> – technická metadata;
- <rightsMD> – administrativní případně legislativní (autorská) práva k objektům, povolení k využívání (licence);
- <sourceMD> – nejčastěji obsahuje popisná metadata analogové předlohy digitálního objektu;
- <digiprovMD> – metadata spojená s digitálními zdroji, proveniencí a jejich životním cyklem (migrace, aktivity, změny, rozhodnutí aj.).

Podobně jako v sekci popisných metadat lze administrativní metadata vložit do každé z výše uvedených částí pomocí <mdWrap> a/nebo linkovat na ně pomocí <mdRef>. Pravidla i atributy jsou stejné. Počet sekcí administrativních metadat v jednom METS záznamu není omezen. Je běžné, že jedna sekce <amdSec> odpovídá např. jedné stránce a obsahuje administrativní metadata pro všechny její reprezentace (např. archivní kopii, uživatelskou kopii a OCR). Je také možné mít pouze jeden výskyt sekce <amdSec> pro celý dokument a v tomto elementu mít vícekrát <techMD> a např. <digiprovMD>, dle počtu událostí a počtu objektů. Jednotlivé části <amdSec> musejí povinně mít ID, které slouží odkazování na administrativní metadata z jiných částí konkrétního METS záznamu.

V návrhu profilu NDK je část administrativních metadat <amdSec> mimo hlavní METS záznam ve vedlejším METS záznamu. Hlavní METS záznam na něj linkuje jako na další objekt. Cílem je zkrácení hlavního METS záznamu a také rychlejší manipulace a využívání samotných záznamů s administrativními metadaty náležejícími vždy k jedné stránce (skenu).

Rada spravující METS standard doporučuje pro administrativní metadata použití několika schémat. První skupinou jsou schémata vyvinutá v Kongresové knihovně za účelem použití v METS záznamech. Mezi ně patří schémata audioMD pro technická metadata audio dokumentů, videoMD pro audiovizuální dokumenty a imageMD pro obrazové dokumenty. Pro metadata původu vzniklo schéma digiprovMD a pro metadata práv schéma rightsMD. Z ostatních schémat, která jsou užívanější než výše jmenovaná, nesmíme opomenout MIX pro záznam technických metadat digitálních obrazových dokumentů a především PREMIS, jehož jednotlivé části nalézají uplatnění jak pro technická (PREMIS Object), tak pro administrativní metadata (PREMIS Events,

PREMIS Agent, PREMIS Rights). Pro metadata práv také existuje jedna z mála specifikací přímo ve schématu METS, a to je METS Rights.

<fileSec>

Tato sekce obsahuje výčet souborů spojených s popisovaným digitálním dokumentem. Může se např. jednat o seznam všech archivních kopií, uživatelských kopií, XML záznamů a OCR souborů, které patří k jednomu konkrétnímu popisovanému svazku monografie. Hlavní element této sekce <fileSec> může mít dceřiné elementy <fileGrp> a <file>. Pomocí <fileGrp> lze soubory slučovat do logických skupin, které definují význam případně použití daného souboru (<fileGrp> pro archivní kopie, <fileGrp> pro uživatelské kopie apod.). Počet elementů <fileGrp> není omezen, vyplývá z potřeb nebo logiky popisu. Z tohoto důvodu je pro <fileGrp> důležitý atribut USE, který drží informaci o typu skupiny souborů (např. master pro archivní soubory, user pro uživatelské, layout pro ALTO XML apod.).

Element <fileGrp> obsahuje jeden nebo více dceřiných elementů <file>, které obsahují odkaz na vlastní soubor na úložišti (většinou cesta ve file systému, může jít také o odkaz mimo systém, např. do prostředí Internetu). Element <file> má mnoho atributů, nejčastěji se používají atributy SEQ (vyjadřuje pořadí ve skupině souborů), SIZE (velikost souboru), MIMETYPE, ID a ADMID nebo DMDID (link na relevantní administrativní nebo popisná metadata). Velmi často můžeme vidět také kontrolní součet v rámci atributů CHECKSUMTYPE a CHECKSUM. Uvedení kontrolních součtů je dalším (volitelným a často využívaným) výskytem tohoto pro dlouhodobou ochranu a správu digitálních objektů důležitého údaje (vedle výskytu v administrativních metadatech).

Odkaz na vlastní soubor lze realizovat dvěma způsoby. První je odkaz na z pohledu METS záznamu externí soubor. Využívá se element <FLocat> (*file location*) podřízený elementu <file>. Link je uveden jako URN, URL, DOI, HANDLE nebo jiný identifikátor externího souboru – viz Ukázka 8.

```
<fileGrp ID="IMGGRP" USE="Images">
    <file ID="IMG00001" ADMID="IMGPARAM00001" MIMETYPE="image/jp2" SEQ="1"
    CHECKSUM="da013954d804aa414cf8401206270b8c" CHECKSUMTYPE="MD5"
    SIZE="9140495">
        <FLocat LOCTYPE="URL" xlink:href="file://./preservation_img/pr-00001.jp2"/>
    </file>
</fileGrp>
```

Ukázka 8 – <fileGrp> METS záznamu (zdroj finská národní knihovna).

Druhou možností jak může být soubor součástí METS záznamu je vnoření binárního kódu souboru přímo do METS <file> elementu. V elementu <file> se použije element <FContent> a v něm dále element <binData>. Data musí být kódována v Base64. Lze vložit také XML podobu dat, element pak není <binData>, ale <xmlData>. Vložení binárního kódu je naprosto nevhodný způsob s ohledem na dlouhodobou ochranu digitálních dat. Neumožní nástrojům provést validaci souboru, charakterizaci, tvorbu technických metadat k digitálnímu objektu apod.

<structMap>

Sekce strukturálních map je jako jediná v METS záznamu povinná (vedle kořenového elementu). Definuje hierarchickou strukturu popisovaného dokumentu a jeho vazby na ostatní sekce METS

dokumentu (např. na administrativní nebo popisná metadata). Je také velmi důležitá pro zpřístupnění a navigaci v popisovaném digitálním dokumentu samotném. Často je díky tomu označována za srdce METS záznamu. Standard METS definuje dva základní typy strukturálních map, fyzickou a logickou, nicméně ponechává možnost definovat vlastní typ strukturální mapy. Fyzická strukturální mapa popisuje fyzickou strukturu dokumentu (v případě digitalizace tedy strukturu předlohy). Taková struktura může být velmi jednoduchá, sled stránek se seznamem reprezentací každé z nich (Ukázka 9 znázorňuje dvě reprezentace – obraz a ALTO XML náležející k první stránce dokumentu). Oproti tomu logická strukturální mapa člení dokument z pohledu struktury do logických celků (kupříkladu kapitoly v knize, obsah aj.).
Hlavní element této sekce, <structMap>, může mít atributy LABEL, TYPE a ID, všechny jsou nepovinné. Atribut TYPE se používá vždy, když je nutno odlišit logickou mapu od fyzické. Ty se totiž často vyskytují současně. Může mít tedy pouze jednu z následujících hodnot – LOGICAL nebo PHYSICAL (viz Ukázka 9).

```
<structMap LABEL="Physical Structure" TYPE="PHYSICAL">
    <div ID="DIVP1" DMDID="MODSMD_PRINT" LABEL="Tyrvään Sanomat" TYPE="Newspaper">
        <div ID="DIVP2" ORDER="1" TYPE="PAGE">
                <fptr>
                    <par>
                        <area FILEID="IMG00001"/>
                        <area FILEID="ALTO00001" BETYPE="IDREF" BEGIN="P1"/>
                    </par>
                </fptr>
        </div>
        …
```

Ukázka 9 – Fyzická strukturální mapa METS záznamu (použit METS záznam finské národní knihovny).

Strukturální mapa, logická i fyzická, sestává z více či méně „zahnízděných“ elementů (ve smyslu *division*, tj. oddělení). Každý může obsahovat devět volitelných atributů. Nejužívanější jsou atributy ID (identifikátor samotného), ORDER (označuje pořadí konkrétního vzhledem k stejného typu nebo vzhledem k celému dokumentu), TYPE (označení typu , úrovně struktury, např. kapitola, číslo, svazek, článek a další), DMDID (reference na popisná metadata; důležité zvláště u úrovní , pro které existuje část popisných metadat, např. článek, kapitola), ADMID (reference na konkrétní část administrativních metadat), LABEL (název konkrétního , tj. oddělení). Pro případ stránky lze využít atributy ORDER a ORDERLABEL následujícím způsobem. ORDER zaznamená fyzické pořadí stránky, ORDERLABEL logické stránkování, tj. jak je na stránce vytištěno (porovnej s řešením v DTD metadatech pro Krameria – viz kapitola 5.3.2.3).
Uvnitř elementu mohou být další dceřiné elementy <mptr> METS Pointer a <fptr> File Pointer. <mptr> je využíván k odkazu za jiný METS dokument. Příkladem využití může být strukturální mapa METS záznamu popisujícího sbírku knih. Každá kniha bude mít svůj vlastní METS záznam, který bude výše uvedeným způsobem linkován z fyzické strukturální mapy. Element <fptr> naproti tomu ukazuje buď přímo na soubor, který reprezentuje a to za pomoci FILEID atributu (viz Ukázka 10 a Ukázka 9), nebo poskytuje více komplexní odkaz do souboru pomocí elementů <area>, <seq> Sequence a <par> Parallel Files. V rámci elementu <area> lze takto

specifikovat začátek a konec té části souboru, která reprezentuje konkrétní , tedy např. článek na stránce (viz Ukázka 11, kde je použit pouze atribut BEGIN a END v ALTO XML náležejícím k článku). Element <par> se používá pro označení (zabalení) více souborů, které mají být zobrazeny zároveň. Ukázka 9 zobrazuje v elementech <par> obrazový soubor a k němu náležející OCR v podobě ALTO XML souboru.

```
<mets:structMap ID="structmap1" TYPE="physical">
    <mets:div TYPE="issue" DMDID="nla.news-issn18339719_19450913">
        <mets:div ID="divpage1" TYPE="page" ORDER="1">
                <mets:fptr FILEID="nlaImageSeq-33385-b.tif"/>
                <mets:fptr FILEID="nlaImageSeq-33385-b.xml"/>
        </mets:div>
        …
```

Ukázka 10 – s vnořeným elementem <fptr>.

```
<div TYPE="PARAGRAPH" ID="DIVL14" ORDER="2">
    <div TYPE="TEXT" ID="DIVL15">
        <fptr> <area FILEID="img0004-alto" BEGIN="P5_TB00004" END="P5_TB00004"/>
        </fptr>
    </div>
</div>
```

Ukázka 11 – Ukázka z logické strukturální mapy, článku s vnořeným elementem <fptr> a v něm element <area> s označením začátku a konce relevantní části souboru.

<structLink>

Sekce strukturálních linků. Hlavní použití této sekce METS spočívá ve vyjádření propojení elementů fyzické a logické strukturní mapy, tedy jakýchkoliv dvou sekcí strukturální mapy. Tato sekce METS byla navržena primárně pro použití pro popis webových stránek. Tento způsob popisu všech částí tvořících webovou stránku v METS záznamu je možný, ale velmi komplikovaný a časově náročný. Proto se tato část METS využívá i pro původní účel velmi zřídka.[278] Linkování je provedeno pomocí elementů <smLink>, které využívají atributy ID sekcí strukturální mapy. První ID je ze sekce ze které link vychází, druhé je do sekce do které link míří [CANTARA, 2005, s. 249]. Tato část METS není v návrhu metadatového profilu pro NDK použita.

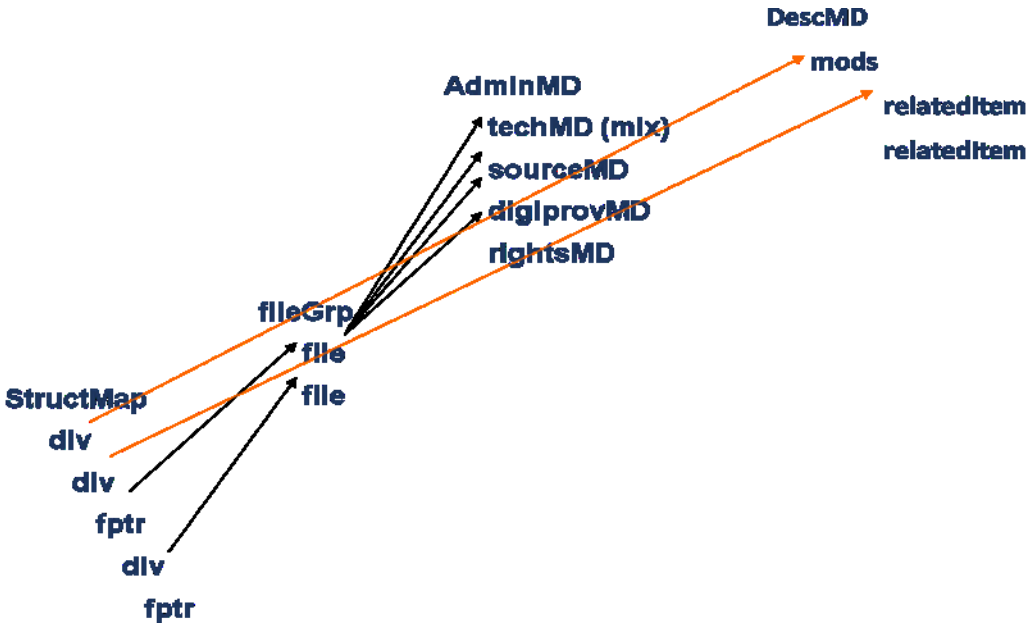
<behaviorSec>

Sekce modelů chování poskytuje informace o spustitelných kódech (aplikacích), které jsou spojeny s obsahem digitálního objektu popisovaného METS záznamem. Příkladem mohou být XSLT šablony pro transformaci XML souboru do HTML podoby pro zpřístupnění; nebo také aplikace na prohlížení digitálních obrazů nebo celých knih; případně obyčejný přehrávač jako je např. QuickTime [CANTARA, 2005, s. 250]. Sekce <behaviourSec> pomocí dceřiného elementu <behaviour> a jeho atributu STRUCTID může linkovat na sekci nebo divizi strukturální mapy METS

[278] V současnosti se do digitálních repozitářů ukládají data vzniklá archivací webu nejčastěji jako ucelené balíky v kontejnerových schématech ARC nebo nověji WARC. Tyto se pak popisují v rámci archivu. Zpřístupnění probíhá nástrojem, který využívá další doprovodné soubory k těmto balíkům, které popisují obsah jednotlivých ARC nebo WARC balíků (např. nástroj WayBack Machine).

záznamu. Podobně pomocí atributu ADMID lze linkovat na administrativní metadata objektu, ke kterému se váží údaje o aplikaci.
Tato sekce METS se běžně nevyužívá. Informace o aplikaci potřebné pro zobrazení popisovaného dokumentu nebo objektu, se ukládají většinou do elementů PREMIS Object záznamu, který může být součástí administrativních metadat METS, zvláště v případech, kdy jde o METS záznam vytvářený pro dlouhodobé uložení v digitálních repozitářích a v LTP systémech. Tato část METS není v návrhu metadatového profilu pro NDK použita.

Jak vypadají výše popsané vztahy mezi částmi METS záznamu ve zjednodušené podobě ukazuje Obrázek 24 níže.



Obrázek 24 – Vztahy jednotlivých hlavních částí METS záznamu [CUNDIFF a TRAIL, 2007].

7.2.1.2 METS profil

K zajištění interoperability a sdílení dat definovala rada METS části tzv. METS profilu a navrhla schéma XML k jeho zápisu, které je nyní ve své druhé verzi[279].
METS profil, podobně jako aplikační profil metadat, je určen k popisu konkrétního METS záznamu pro konkrétní typ dokumentu. Pomáhá jak autorům, tak programátorům vytvořit METS záznam, který konkrétnímu profilu odpovídá. Profil může předepisovat použití konkrétních schémat v METS záznamu, kontrolovaných slovníků, specifikovat uspořádání a využití elementů a jejich atributů pro popis konkrétního typu dokumentu. Může také doporučovat nástroje na vytváření METS záznamů, které konkrétnímu profilu odpovídají [McDONOUGH, 2003]. Jednotlivé profily je možno dobrovolně registrovat na stránkách schématu METS[280]. Profily k registraci musí mít konkrétní formální podobu, jak ji specifikuje stránka METS určená pro profily[281] a již zmíněné XML schéma.

[279] http://www.loc.gov/standards/mets/profile_docs/mets.profile.v2-0.xsd
[280] http://www.loc.gov/standards/mets/mets-profiles.html
[281] http://www.loc.gov/standards/mets/profile_docs/components.html

7.2.1.3 PREMIS v METS záznamu

PREMIS specifikace je vytvářena s cílem jejího použití v kontejnerovém schématu METS, přesněji v jeho části pro administrativní metadata <amdSec>, viz kapitola 7.2.1.1. Oba standardy se snaží si vyjít vstříc, vznikaly a jsou upravovány s tím, že tato symbióza je velmi podstatná. První možností je vložit celý záznam PREMIS do elementu <amdSec>, nejčastěji se k tomuto účelu používá část <digiprovMD>. Vložený záznam PREMIS musí být uveden elementem <premis>. Obě specifikace to podporují, nejde ale o časté řešení. Častěji jsou jednotlivé části PREMIS, vyplývající z jeho datového modelu, použity v dceřiných elementech <amdSec>. Existuje doporučení řídícího výboru METS, které toto přesně specifikuje – viz Tabulka 5.

část PREMIS	**část METS <amdSec>**	**poznámka**
PREMIS Object	<techMD>	nejčastější a doporučovaná varianta pro popis souboru nebo bitstreamu
	<digiprovMD>	v případě potřeby uvést zde technické informace o reprezentaci, nepoužívá se často
PREMIS Event	<digiprovMD>	
PREMIS Rights	<rightsMD>	
PREMIS Agent	<digiprovMD>	pokud je Agent uveden v souvislosti s Událostí (eventem)
	<rightsMD>	pokud je Agent uveden v souvislosti s deklarací práv

Tabulka 5 – možnosti vložení PREMIS do záznamu METS (část <amdSec>).

Pokud je využívána část PREMIS standardu PREMIS Agent v souvislosti s nějakou *Událostí* nebo *Právy* (více o těchto entitách viz kapitola 7.2.3.2), měly by tyto být ve vlastní části <digiprovMD> nebo <rightsMD>, minimalizuje to zbytečné opakování (stejný *Agent* může být spojen s jinými *Událostmi* a *Právy*). METS také specifikuje sekci <sourceMD>, kterou Tabulka 5 neuvádí. Je to proto, že tato sekce je určena pro informaci o fyzické předloze, není tedy relevantní pro PREMIS metadata, která jsou o digitálním objektu. Do sekce <sourceMD> se vkládá např. původní bibliografický záznam předlohy (MODS, MARCXML).

Mezi standardy METS a PREMIS je v některých elementech, hlavně u technických metadat, překryv – viz Tabulka 6. Existuje několik možností, jak se s tím vypořádat. První z nich je technická metadata uvádět pouze v METS záznamu a rezignovat na ně v záznamu PREMIS. Tato možnost ovšem vylučuje jednoduché samostatné použití PREMIS záznamu, kterému by chyběly podstatné technické informace. Druhá možnost je uvést technická metadata pouze v PREMIS záznamu. Poslední možností je využít oba způsoby. Právě třetí možnost je nejvíce využívána. Systémy zpracovávající metadata METS vědí, v jakém kontextu jsou jednotlivé elementy uvedeny dle jmenného prostoru, nepůsobí to tedy žádný problém.

METS	**PREMIS**
METS: SIZE	PREMIS: size
METS: CHECKSUM, CHECKSUMTYPE	PREMIS: fixity

METS: MIMETYPE	PREMIS: format
METS ID/Idref: používá se ke spojení metadat napříč sekcemi METS záznamu	PREMIS identifiers: linkování mezi různými typy entit
METS structMap: strukturální vztahy linkující elementy strukturální mapy na jednotlivé objekty a jejich metadata	PREMIS <relationship>: může vyjádřit různé druhy vztahů, včetně strukturálních

Tabulka 6 – Ukázka překrývání elementů METS a PREMIS.

7.2.2 Popisná metadata

Pro popisná metadata v profilu NDK byl zvolen standard MODS, který je dnes v podobných projektech běžný. Výhodou je, že má silnou podporu komunity, vznikl pro popis digitálních objektů vzniklých převodem z analogové podoby, takže dokáže popsat i fyzickou předlohu. Ve prospěch standardu hraje také jeho flexibilita a možnost hierarchizace záznamů.
V prostředí NK ČR se začal používat standard MODS poprvé ne přímo pro popis intelektuálních entit jak je chápe datový model PREMIS, ale byl využit jako výměnný formát pro OAI-PMH protokol v digitální knihovně Manuscriptorium (rok 2005). MODS se díky slovnímu vyjádření polí a své jednoduchosti pro OAI-PMH hodí lépe než MARC21 a v případech, kdy chceme posílat více údajů, je vhodnější než minimalistický formát Dublin Core. První použití jako popisného standardu nastane ovšem až nyní v projektu NDK. Specifikace obsahuje popis monografií a periodik a jejich jednotlivých úrovní. MODS standard má 19 elementů vrchní úrovně a dva kořenové elementy. Každý element má dceřiné elementy a velký počet atributů. Elementy se mohou libovolně opakovat a nemusí být za sebou v určitém pořadí. Je tak jednoduché vytvářet potřebné hierarchie. Popis vývoje a vlastností standardů Dublin Core a MODS není uveden zde, ale v kapitole 4.8.

Popisná metadata využívaná v repozitářích a LTP systémech jsou různá. Někdy je to pouze minimální počet údajů uložených s cílem zajištění vyhledávání v rámci systému. V takovém případě je kompletní sada popisných metadat dostupná vně LTP systému, např. v katalogu. Mnoho institucí nevidí důvod, proč opakovat údaje z katalogu v repozitáři nebo digitální knihovně, stačí je pouze vhodně propojit. Výhodou je, že při změně údajů v katalogu není potřeba měnit metadata v LTP systému. Katalogizační záznam a odpovídající entita v LTP systému jsou propojeny identifikátorem. K tomuto přístupu stačí např. Dublin Core záznam v repozitáři (LTP systému) uložený.[282] Pokud chceme ukládat kompletní a rozsáhlá popisná metadata, používá se pro digitalizované monografie a periodika nejčastěji právě MODS, případně další standardy pro různé typy digitálních objektů (VRA, EAD). Výhodou je, že intelektuální entita z hlediska popisu je v LTP kompletní a nezávislá. Touto cestou se na základě předchozích postupů v digitalizaci rozhodla jít NK ČR, proto profil pro NDK obsahuje poměrně rozsáhlou specifikaci struktur a elementů MODS. Minimální záznam MODS je založen na polích minimálního záznamu monografií a periodik v MARC21, jak jej stanoví NK ČR. Dublin Core uvedený ve specifikaci metadat pro NDK je součástí balíčku metadat z digitalizace z toho důvodu, aby nebylo nutné vytvářet Dublin Core z MODS *on-*

[282] Tento přístup je preferován LTP systémem Rosetta od firmy Ex Libris. V něm je pouze pár polí Dublin Core, které slouží vyhledávání dle názvu apod., každá entita obsahuje link do katalogu přes tzv. CMS ID. Kompletní popisná metadata lze do systému vložit jako další digitální objekt, tedy reprezentaci intelektuální entity.

the-fly v případě potřeby (např. pro OAI-PMH). Dublin Core se také hodí pro aplikace zpřístupnění, které s ním již v současnosti pracují (Kramerius4).

Specifikace MODS pro NDK se od předchozích struktur popisných metadat pro novodobé dokumenty podstatně liší, a to tím, že neexistuje entita ročník u periodik. Důvodem je fakt, že jde o uměle vytvořenou entitu, která obsahovala vždy minimum údajů. Tyto údaje (číslo ročníku, rok vydávání) jsou v MODS specifikaci pro NDK součástí popisu každého čísla. Pokud tedy uživatel zadá ve vyhledávání např. výraz „Rudé právo 1979“, dostane seznam všech čísel toho roku na základě metadat čísla. Pokud bude aplikace zpřístupnění k tomuto konceptu uzpůsobená, uživatel nemusí vůbec poznat, že ročník jako samostatná entita neexistuje. Tento postup je ve světových knihovnách běžný. Digitalizují se jednotlivá čísla periodika. Aplikace zpřístupnění pak vytvářejí agregované zobrazení čísel po měsících, letech apod., např. v podobě kalendáře.

7.2.3 PREMIS

Standard PREMIS byl jasnou volbou pro vyjádření administrativních, ochranných a částečně technických metadat. Pro ochranná metadata jde o zdaleka nejrozšířenější standard, na jehož datovém modelu staví další projekty, architektura procesů LTP systémů a repozitářů i obecné koncepty práce s digitálními objekty a jejich hierarchiemi. První nasazení PREMIS proběhlo v NK ČR v roce 2008 pro ochranná a technická metadata objektů vzniklých digitalizací novodobých dokumentů. Tehdy byla implementována pouze část standardu PREMIS Object, která je schopna popsat základní technická metadata pro jakýkoliv digitální objekt a také ochranná metadata pro digitální objekt [HUTAŘ, 2008a]. Nový návrh pro projekt NDK vychází mj. ze zkušeností, získaných implementací z roku 2008. Od roku 2008 se standard PREMIS ještě více rozšířil a je základem i komerčních řešení LTP systémů Rosetta a SDB. Je určen pro zapouzdření do METS záznamu, případně pro samostatné použití. V NDK digitalizaci bude PREMIS použit uvnitř vedlejšího METS záznamu.

Jak bylo již uvedeno v kapitole 4.3.2.2, je těžké vytyčit jasné hranice v tom, jaké typy informací spadají do sféry ochranných metadat. Obecný konsensus platí pro pět oblastí informací relevantních pro ochranná metadata [LAVOIE a GARTNER, 2005, s. 5]:

- ***provenience:*** ochranná metadata zaznamenávají informaci o historii digitálního objektu;
- ***autenticita:*** ochranná metadata obsahují informaci dostatečnou k ověření toho, že archivovaný digitální objekt je tím, za co se vydává, nebyl změněn, ať záměrně nebo nechtěně, aniž by o tom byl záznam v metadatech;
- ***ochranné aktivity:*** ochranná metadata dokumentují události provedené na objektech v rámci jejich ochrany včetně jejich dopadu na vzhled, vlastnosti a funkcionalitu objektu;
- ***technické prostředí:*** ochranná metadata popisují technické nároky na HW a SW (operační systém a aplikace) potřebné k zobrazení a použití digitálního objektu ve stavu, v jakém je aktuálně uložen v archivačním systému;
- ***management práv:*** ochranná metadata zaznamenávají jakákoliv omezení vyplývající z autorského zákona, která by mohla ohrozit schopnost systému provést ochranná opatření na digitálním objektu a poskytnout objekt současným a budoucím uživatelům.

V profilu metadat pro NDK se používají již všechny části standardu, tedy PREMIS Object (Objekt), Event (Událost), Agent (Agent), kromě části Rights (Metadata práv). Rozhodnutí bylo takové, že údaje o právech, které se váží k digitálním objektům, budou součástí aplikace zpřístupnění. Z pohledu dlouhodobé ochrany by ovšem bylo vhodné metadata práv do specifikace zapracovat,

není to ale nutné ve specifikaci metadat pro digitalizaci – více viz kapitola 7.2.3.2. Specifikace profilu metadat pro NDK je vytvořena pro tvorbu metadat v digitalizaci, PREMIS tedy v profilu NDK pokrývá okamžik vytvoření objektu a jeho úpravy. Použití těchto metadat v LTP systému je dalším krokem, který specifikace neřeší. METS záznam není specifikován tak, že by bylo možné jej bez jakékoliv změny použít v LTP systému. Na vstupu se musí přidat další metadata, záleží ovšem jak s PREMISem pracuje samotný LTP systém. Profil NDK výše popsaným nárokům na další procesy v LTP systému maximálně napomáhá, a to právě zaznamenáním a popsáním procesů provedených v digitalizaci. Ochranná metadata PREMIS jsou pro LTP systém vstupní informací.

Vývoj standardů pro ochranná metadata probíhal od roku 2000. Aktivní byly národní knihovny Austrálie, Nového Zélandu aj. – viz kapitola 4.8.2. Návrhy vycházely z podobných potřeb a velmi často jeden návrh rozpracovával návrh starší. V červnu 2002 vznikla tzv. *Preservation Metadata Framework* pracovní skupina organizací OCLC a RLG. Vytvořila komplexní, high-level popis typů informací, které by měly být součástí ochranných metadat. Jako výchozí bod byl využit referenční rámec OAIS. Později vznikla mezinárodní pracovní skupina PREMIS (*Preservation Metadata: Implementation Strategies*), jejíž členové jsou z oblasti knihoven, akademické oblasti, oblasti muzeí, archivů, vládních institucí i komerčního sektoru. Vedoucími osobami pracovní skupiny od počátku byly a dodnes jsou Priscilla Caplan a Rebecca Guenther. Cíly pracovní skupiny bylo mj. [OCLC/RLG PREMIS WORKING GROUP, 2004, s. 9]:

- definovat lehce implementovatelnou sadu „klíčových“ elementů, které by byly široce použitelné uvnitř komunity zabývající se ochranou digitálních dat;
- vytvořit datový slovník pro podporu sady elementů ochranných metadat;
- vyzkoušet a ohodnotit alternativní strategie pro zápis, uložení a správu ochranných metadat uvnitř digitálních repozitářů, jakož i výměnu metadat mezi nimi;
- vytvořit strategie pro zápis, balení, uložení, správu a výměnu ochranných metadat.

Jedním z prvních výstupů pracovní skupiny byla zpráva *Implementing Preservation Repositories for Digital Materials: Current Practice And Emerging Trends In The Cultural Heritage Community* [OCLC/RLG PREMIS WORKING GROUP, 2004]. O rok později, tedy v roce 2005, již následoval PREMIS datový slovník v první verzi.

PREMIS standard byl vyvinut s cílem vytvoření použitelné sady základních elementů metadat pro ochranu digitálních dokumentů, která by byla široce implementovatelná v komunitě zabývající se dlouhodobou ochranou digitálních dat. PREMIS se nezabývá popisnými metadaty, tam přenechává místo již zavedeným standardům. Nespecifikuje speciální technická metadata jmenovitě pro různé typy digitálních objektů, ale soustředí se na klíčové společné elementy (základní množinu informací vhodných pro všechny typy archivů a objektů). V NDK digitalizaci bude PREMIS Object použit pro popis obrazových dat i dat textových (TXT, XML). Pokud chce instituce používat dodatečná technická metadata, má vždy možnost ještě navíc využít jiné standardy relevantní pro daný typ digitálního objektu. Podobné je to s dokumentací k HW a SW potřebných pro zobrazení a použití digitálního objektu. PREMIS obsahuje základní informace o HW a SW potřebném pro zobrazení objektu, zbytek přenechává např. registrům formátů. Použití PREMIS nepředpokládá specifický systém, technologii, systémovou a databázovou architekturu nebo ochrannou strategii.

PREMIS standard není řešením *out-of-the-box*, výběr elementů je veden potřebou instituce a cílem ochranných metadat a musí být ustanoven v konkrétním aplikačním profilu metadat.

V případě profilu pro digitalizaci NDK je cílem zaznamenat procesy v digitalizaci. PREMIS např. neobsahuje informace o metadatovém záznamu samotném (kdo záznam vytvořil, kdy byl změněn aj.) [DAPPERT, 2009]. Metadata práv jsou omezena pouze na informace relevantní pro procesy v repozitáři.

Ve verzi PREMIS 1.0 byla vytvořena XSD schémata jednotlivě pro všechny části datového modelu PREMIS (Object, Agent, Event, Rights). Povinné bylo použití PREMIS Object, ostatní části byly volitelné. Od verze 2.0 existuje pouze jedno XSD schéma, které kombinuje dohromady všechna schémata z verze 1.0. I tak lze ovšem, stejně jako ve verzi 1.0, použít jednotlivé části odděleně, díky tomu, že každá z nich je definována globálně. Takto se s XSD schématem pracuje i v profilu pro NDK, kde jednotlivé části PREMIS jsou určeny pro vložení do METS záznamu a to od sebe odděleně.

7.2.3.1 Datový slovník PREMIS

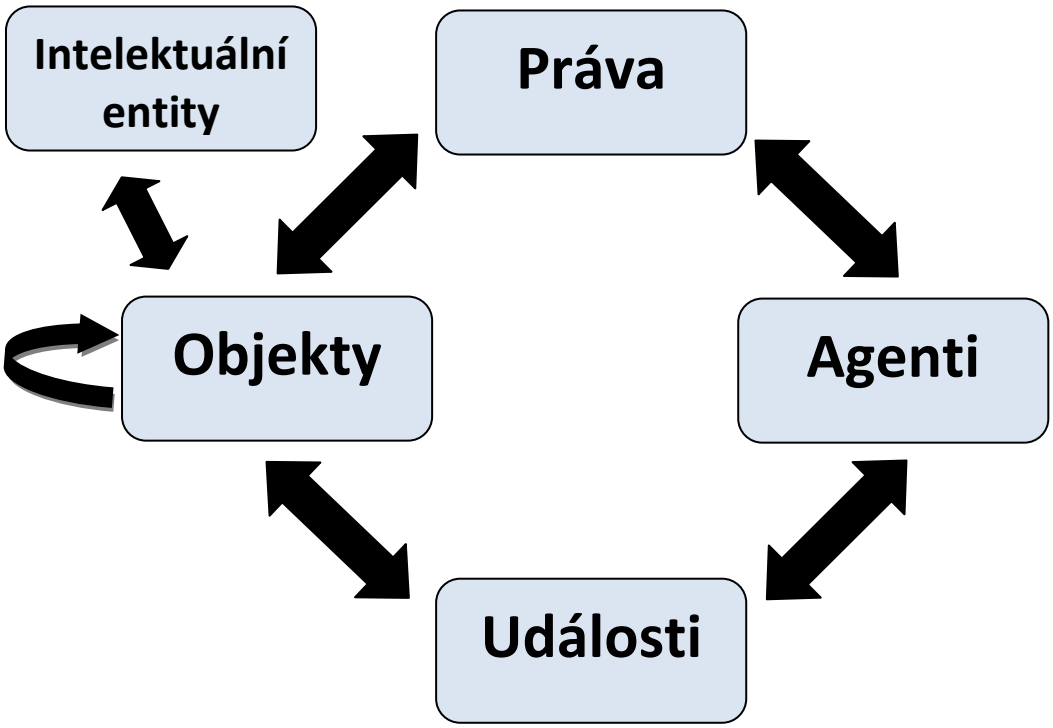
Hlavním dokumentem standardu PREMIS je vedle XSD schématu *Datový slovník PREMIS* (*PREMIS Data Dictionary*). Ten popisuje obecný datový model pro organizování a vlastně uvažování o ochranných metadatech. Dává také velmi dobrý návod na jejich konkrétní implementaci. Datový slovník sestává z popisu datového modelu PREMIS, ze seznamu jednotlivých elementů (sémantických jednotek). Každý element je pak velmi podrobně popsán z hlediska použití, sémantiky a doplněn příklady možných hodnot plnění. Datový slovník PREMIS staví na referenčním modelu OAIS, který poskytuje koncepční základy specifikací typů informačních objektů a balíčků pro archivované digitální objekty a také strukturu s nimi souvisejících metadat. Musíme ovšem zmínit, že PREMIS a OAIS používají určité termíny (terminologii) odlišně. Rozdíly obyčejně odrážejí fakt, že sémantické jednotky PREMISu potřebují více specifikovat než definice v rámci OAIS, což se ale předpokládá, protože se v podstatě jedná o rozdíl mezi obecným rámcem a konkrétní implementací do něj [OCLC/RLG PREMIS WORKING GROUP, 2005, s. ix].
V roce 2008 vyšla nová upravená verze datového slovníku 2.0. Opravy původní verze 1.0 z roku 2005 byly pečlivě plánovány a projednávány. Změny, a to platí i pro další verze, vždy vycházejí z komentářů a potřeb komunity používající ochranná metadata.[283] Jedním z významných doplnění bylo rozpracování části, která se věnuje metadatům práv. V nové verzi je problematika rozvedena více než ve verzi 1.0. Jako v původní verzi jsou metadat práv v PREMIS 2.0 určena primárně k podpoře automatických procesů, které určí, zda ochrannou akci na konkrétním digitálním objektu je možno provést vzhledem k povolením a autorským právům. Ve verzi 2.0 je původní element <permissionStatement> nahrazen kontejnerovým elementem <rightsStatement>. Ten může popsat tři formy práv intelektuálního vlastnictví (copyright, licenci a statut). Metadata práv obsahují mj. údaje o *Agentech* (osoba, systém nebo instituce), kteří práva spravují, o *Objektech*, kterých se práva dotýkají.
Změnu zaznamenaly také sémantické jednotky popisující signifikantní vlastnosti. V původní verzi PREMIS byla k jejich popisu určena pouze jedna sémantická jednotka. Byla nestrukturovaná a neměla přesná pravidla použití. Bylo možné vyplnit slovní popis každé vlastnosti, nebo např. pouze jejich kódy. Ve verzi PREMIS 2.0 je možno vyjádřit signifikantní vlastnosti elegantněji.

[283] Návrh na změnu v Datovém slovníku PREMIS má formální podobu, zasílá se řídícímu výboru, je přijat a během 2 měsíců diskutován komunitou, následně je návrh probírán minimálně na dvou setkáních výboru, rozhodnutí je pak vystaveno na pracovní wiki a objeví se v další verzi slovníku [CAPLAN a ZWAARD, 2011].

Sémantická jednotka byla nahrazena dvěma, z nichž první uvádí obecnou skupinu signifikantních vlastností *Objektu* (obsah, funkčnost, vizuální podoba aj.). Druhá sémantická jednotka popisuje již konkrétní signifikantní vlastnost náležející do obecné množiny.
Podstatným rozšířením standardu PREMIS bylo přidání možnosti obohatit záznam o vlastní XML metadata. To je vhodné v případech, kdy instituce potřebuje rozšířit konkrétní údaje nebo podrobnější schéma. V původní verzi toto bylo problematické, schéma neobsahovalo mechanismus vložení jiného XML záznamu do záznamu PREMIS. Nová verze tento mechanismus nabízí v sedmi sémantických jednotkách (elementech): <significantProperties>, <objectCharacteristics>, <creatingApplication>, <environment>, <signatureInformation>, <eventOutcomeDetail> a <rights>. Bylo rozhodnuto, že tyto sémantické jednotky jsou těmi, u kterých je největší pravděpodobnost, že by uživatelé PREMISu mohli mít potřebu jejich popis rozšířit.[284] Zatím poslední verzí datového slovníku PREMIS je verze 2.1 z ledna 2011 – viz [PREMIS EDITORIAL COMMITTEE, 2011]. Nejpodstatnější změnou oproti verzi 2.0, kromě upřesnění stávajících elementů a drobných úprav, je rozšíření údajů o *Agentovi*, např. o elementy <agentNote> a <agentExtension>.

7.2.3.2 Datový model PREMIS

Pro ulehčení pochopení celkové logické organizace metadatových elementů PREMIS vyvinula pracovní skupina PREMIS jednoduchý model pěti typů entit. Jsou jimi: *Intelektuální entity*; *Objekty*; *Události*; *Práva* a *Agenti* – viz Obrázek 25. Entity jsou oblasti, které jsou relevantní pro dlouhodobou ochranu digitálního objektu, který chceme popsat pomocí standardu PREMIS.



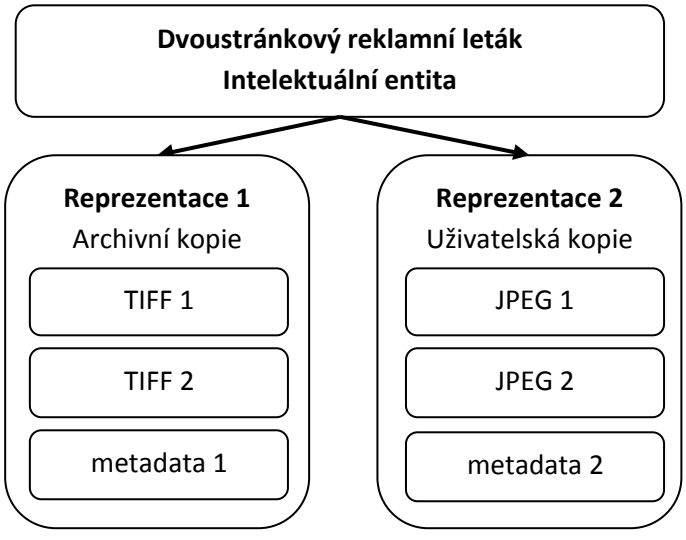
Obrázek 25 – Datový model PREMIS.

[284] K dispozici je např. schéma copyrighMD, vytvořené v *California Digital Library*. To by mohlo být vhodné na rozšíření metadat o právech PREMIS.

Na znázornění datového modelu jsou entity znázorněny jako boxy, mezi kterými jsou vykresleny vztahy se šipkami. Jejich směr ukazuje směr vztahových vazeb definovaných v PREMIS datovém slovníku. Např. šipka od entity *Práva* směrem k *Agentům* znamená to, že metadata definovaná pro *Práva* obsahují sémantické jednotky k identifikaci souvisejících *Agentů*. Oboustranná šipka označuje skutečnost, že jsou definovány oba vzájemné vztahy.
Popis jednotlivých entit níže částečně vychází z workshopu Priscilly Caplan [CAPLAN a ZWAARD, 2011] a z její prezentace ze školení PREMIS [CAPLAN, DAPPERT a ENDERS, 2010], které probíhá každý rok při příležitosti relevantní velké konference (např. iPRES, Archiving apod.).

Intelektuální entita

Intelektuální entita je podle datového modelu PREMIS verze 2.0 [PREMIS EDITORIAL COMMITTEE, 2011, s. 6] množina informací, která je považována za jednu intelektuální jednotku pro účely správy a popisu, např. konkrétní kniha, mapa, databáze aj. Jedna *Intelektuální entita* může obsahovat jiné *Intelektuální entity* (např. fiktivní monografie „Dějiny města Horní Dolní" má 2 svazky; nebo webové sídlo obsahuje jednotlivé webové stránky).
Intelektuální entita má jednu nebo více digitálních reprezentací – viz Obrázek 26. Digitální reprezentace je podle datového slovníku PREMIS [PREMIS EDITORIAL COMMITTEE, 2011, s. 7] množina všech počítačových souborů, které jsou potřebné k tomu, aby je mohl SW reprodukovat jako jednu *Intelektuální entitu* – např. 32 souborů ve formátu JPEG2000, tedy 32 digitalizovaných stránek tištěné knihy a k nim náležející soubor s metadaty „reprezentují" konkrétní podobu Máchova Máje (nakladatel J. L. Kober, 1926). Další reprezentací stejné entity by byla např. digitální kopie v jednom PDF souboru.



Obrázek 26 – Intelektuální entita dle PREMIS a její reprezentace.

Objekt

Objekt nebo digitální objekt je samostatná (diskrétní) jednotka informace v digitální formě. Tato informace tvoří data (tj. digitální objekty) a je tedy tím, co repozitář vlastně ukládá. PREMIS rozlišuje tři typy *Objektů*, ke kterým mohou vznikat ochranná metadata. Jde o:

- ***soubor (file)*** – pojmenovaná sekvence bytů v určité struktuře, která je známá operačnímu systému, např. jeden soubor PDF;
- ***reprezentaci (representation)*** – sada souborů včetně metadat, kterou lze zobrazit jako kompletní *Intelektuální entitu*; např. dva soubory JPEG, které tvoří dohromady dvě stránky letáku (viz Obrázek 26) a soubor s metadaty tvoří dohromady reprezentaci *Intelektuální entity*;
- ***bitstream*** – data uvnitř souboru s vlastnostmi vhodnými pro ochranné aktivity (musí mít strukturu nebo formátování aby mohl existovat jako samostatný soubor); např. soubor TIFF obsahuje hlavičku a jednu naskenovanou stránku, tj. jeden bitstream s vlastní sadou sémantických jednotek; samostatný bitstream jsou také např. metadata vložená do hlavičky souboru JPEG2000.

Repozitář může provádět správu buď pouze pro soubory, nebo i pro reprezentace nebo bitstream. Pokud je správa prováděna na úrovni reprezentace, lze vyjádřit sémantické jednotky (elementy), které jsou určeny pro reprezentace; pokud je správa také na úrovni souborů, musí se pomocí PREMISu vyjádřit vlastnosti náležející k souborům a zároveň je nutné vyjádřit vztahy mezi soubory a reprezentacemi. To samé platí také pro bitstream.

Událost

Činnost nebo proces, který zahrnuje nejméně jeden *Objekt* nebo *Agenta*, kteří jsou známi repozitáři (repozitář o nich „ví“). Pomocí *Událostí* PREMIS dokumentuje provenienci, tj. zaznamenává historii *Objektu* během jeho existence, od jeho vzniku po jeho uložení. Každý repozitář eviduje a do metadat ukládá pouze omezenou množinu typů *Událostí*. Vybrat typy *Událostí*, o nichž je potřeba uchovat záznam spolu s digitálním objektem, je záležitostí administrátora repozitáře. Ten musí určit, jaké *Události* to budou a jaká míra podrobností o nich se bude uchovávat (např. název *Události*; *Agent*, který ji provedl; důvod provedení; výsledek *Události* apod.). *Události* musí vznikat a do metadat se zaznamenávat již v digitalizaci (např. pro procesy skenování, ořezy, narovnání aj.). K tomu je uzpůsoben navrhovaný profil pro NDK. Další nové *Události* přibývají v repozitáři během uložení digitálního objektu. Do repozitáře tedy již přicházejí digitální objekty, v jejichž metadatech jsou zaznamenány *Události* z procesu digitalizace. Níže v Tabulce 7 jsou na ukázku uvedeny typy *Událostí* z procesů digitalizace i z repozitáře, které uvádí Angela Dappert [DAPPERT, 2009, snímek 89].

Vytvoření (skenování)	Komprese
Validace	Dekomprese
Antivirová kontrola	Vymazání
Validace digitálního podpisu	Zpřístupnění
Ověření kontrolního součtu	Vstup do repozitáře (*ingest*)
Vytvoření kontrolního součtu	Migrace
Normalizace	Replikace

Tabulka 7 – ukázka typů Událostí, převzato z [DAPPERT, 2009, snímek 89].

Agent

Osoba, organizace nebo SW aplikace spojený s *Událostmi* nebo *Právy* vážícími se k digitálnímu *Objektu* (souboru, reprezentaci, bitstreamu). *Agenti* jsou spojeni s *Objekty* nepřímo přes *Události* a *Práva* – viz datový model PREMIS, Obrázek 25. *Agenti* nejsou v datovém slovníku do detailu specifikováni; nejsou považováni za nejdůležitější část ochranných metadat. Datový slovník navíc umožňuje vložení jiného standardu metadat o *Agentech* přímo do PREMIS záznamu, pomocí elementu <extension>. Příkladem *Agentů* mohou být: Priscilla Caplan (osoba); NK ČR (organizace); datový repozitář NK ČR nebo NUŠL (systém), JHOVE verze 2.0 (SW). Nejčastějším použitím je název SW aplikace, která provádí konkrétní *Událost* nebo osoby, instituce, firmy, které provádí další *Události*.

Práva

Deklarace práv ve smyslu udělení jednoho nebo více práv nebo povolení náležejících k *Objektu* a/nebo *Agentovi*. PREMIS není komplexní nástroj na vyjádření práv, může zaznamenat pouze povolení typu „Agent X poskytuje povolení Y repozitáři ohledně Objektu Z“. *Práva* jsou nejčastěji spojena s licencí, smlouvou nutnou pro provádění konkrétních *Událostí* v repozitáři. Zaznamenávání autorských práv a typů licencí může být aktuální např. pro dokumenty přicházející do knihovny v elektronické podobě, který může mít různá omezení a lhůty. Příkladem vyjádření *Práv* může být např.: „Jan Novák poskytuje repozitáři NUŠL povolení k vytvoření 3 kopií souboru moje_disertace.pdf pro účely ochrany“. I z tohoto důvodu metadata práv nejsou součástí profilu metadat pro NDK, nejde totiž o návrh metadatového profilu pro LTP systém, ale pro digitalizaci. Návrh profilu vznikal s vědomím, že funkční LTP systém bude mít možnost vyjádřit práva v rámci systému. Podobně je to řešeno v aplikaci zpřístupnění Kramerius. Zaznamenávat údaje o právech pro digitalizované věci v projektu NDK, u kterých je jasné, kdy vyprší autorská práva podle roku vydání, není rozumné řešení. V případě, že bychom ke každému objektu nebo reprezentaci zaznamenali, že autorská práva vyprší v konkrétním roce, který je závislý na aktuálním autorském zákoně, museli bychom v případě změny tohoto zákona měnit všechny údaje v metadatech práv v celém repozitáři. U digital-born dokumentů to bude samozřejmě jiné.

Datový model PREMIS také popisuje vztahy mezi entitami a jejich vlastnosti, které jsou vyjádřeny pomocí sémantických jednotek (*semantic units*), tedy vlastně jednotlivých elementů metadat. Sémantická jednotka je „vlastností“ entity a vyjadřuje to, co potřebujeme k dlouhodobé ochraně digitálních dat vědět o *Objektu*, *Události*, *Agentovi* a/nebo *Právech*. Rozlišujeme dva druhy sémantických jednotek:

- ***kontejner*** – seskupuje dohromady více sémantických jednotek (v záznamu PREMIS zapsaného v XML odpovídá kontejner elementu vrchní úrovně, který obsahuje dceřiné elementy);
- ***sémantická část*** – sémantické jednotky seskupené pod stejným kontejnerem (odpovídají v záznamu PREMIS v XML podobě dceřiným elementům).

Příkladem může být zápis v Ukázce 12:

ObjectIdentifier [kontejner]
    ObjectIdentifierType [sémantická část]
    ObjectIdentifierValue [sémantická část]

Ukázka 12 – Zápis sémantické jednotky PREMIS.

Datový model PREMIS podporuje vyjádření vztahů mezi entitami a také mezi jednotlivými typy *Objektů* na stejné nebo různé úrovni [DAPPERT, 2009, snímek 93]. Vztah může být mezi částmi celku (*Intelektuální entity*), pak je takový vztah strukturální (A je součástí B). Pokud jde o vztah mezi originální reprezentací a novou reprezentací vzniklou migrací, můžeme mluvit o vztahu odvozeném (A je naskenováno z B; A je verzí B apod.). Vztahy mezi jednotlivými entitami lze vyjádřit identifikátory, které jsou obsaženy v tzv. linkovacích sémantických jednotkách, které každá entita obsahuje. *Objekt* tak například odkazuje na konkrétní *Událost* v záznamu PREMIS pomocí identifikátoru *Události* – viz Ukázka 13 níže.

relationship [část popisu Objektu A]
relationshipType = derivation
relationshipSubType = is source of
relatedObjectIdentification [identifikátor Objektu B]
relatedObjectIdentifierType = repositoryID
relatedObjectIdentifierValue = F004400
relatedEventIdentification [identifikátor události – migrace]
relatedEventIdentifierType = repEventID
relatedEventIdentifierValue = E0192

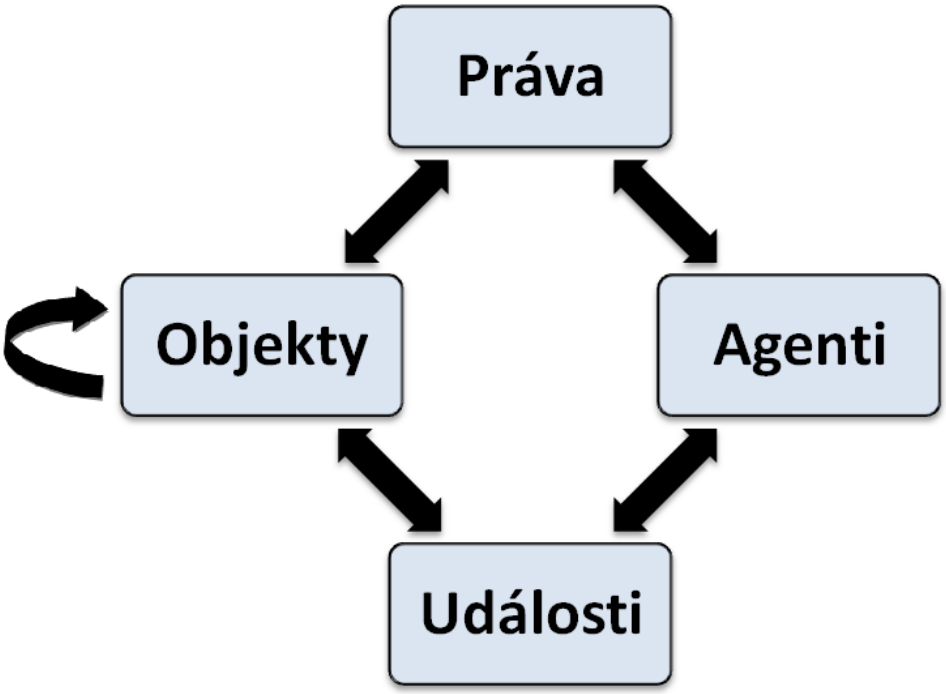
Ukázka 13 – Vztahy mezi sémantickými jednotkami PREMIS.

7.2.3.3 Změny datového modelu PREMIS a výhled pro verzi 3.0

Datový model od své první verze prošel několika úpravami, které byly spíše menšího rozsahu. Největší skok proběhl mezi verzemi 1.0 a 2.0 (současná je 2.1). Větší změnou byla úprava vztahů mezi entitami. Ty jsou dokumentovány pomocí metadat spojených s jednotlivými entitami, např. vztah mezi *Objektem* a *Událostí* lze zaznamenat pomocí sémantické jednotky <linkingEventIdentifier>, která je součástí *Objektu*. V datovém modelu verze 1.0 byla většina vztahů mezi entitami dvojsměrných (metadata *Objektu* obsahovala údaje o související *Události* a naopak). Existovaly ale i vztahy jednosměrné, což v nové verzi datového modelu již nelze. Jednosměrné vztahy byly mezi *Událostí* a *Agentem* a mezi *Právy* a *Agentem*. Vztah *Agenta* a *Práv* bylo možné vyjádřit pouze v záznamu *Práv*, vztah *Události* a *Agenta* pouze v záznamu *Události*. Díky tomu, že PREMIS 2.0 má všechny vztahy mezi entitami dvousměrné, je datový model daleko jednodušší a také flexibilnější.

Markantnější změny datového modelu jsou plánovány pro verzi 3.0. Mluvilo se o nich již 16. května 2011 na workshopu *Implementing PREMIS to support digital preservation*, který byl součástí konference Archiving 2011 v Salt Lake City. Verze 3.0 by měla být publikována počátkem roku 2012. Workshop vedla jedna z hlavních postav vývoje PREMIS a ochranných metadat vůbec, Priscilla Caplan. Hlavní změnou bude přepracování stávajícího datového modelu PREMIS – viz Obrázek 25. Tzv. *Intelektuální entita*, která je součástí datového modelu, ale nebyla popisována v rámci PREMISu a tedy ani datovém slovníku, bude nově dalším typem *Objektu*, který lze pomocí standardu PREMIS popsat. Vedle stávajících typů *Objektů*: soubor, reprezentace a bitstream bude tedy možné popis PREMIS vztáhnout také k *Intelektuální entitě*. Většina elementů vztahujících se v současném modelu k reprezentaci se tak bude vztahovat i k *Intelektuální entitě*, samozřejmě

např. kromě elementů, vyjadřujících technické prostředí pro zobrazení reprezentace [CAPLAN a ZWAARD, 2011]. Nový datový model znázorňuje Obrázek 27.



Obrázek 27 – Návrh nového datového modelu PREMIS [CAPLAN a ZWAARD, 2011].

V nové verzi bude také podstatně rozšířena možnost popisu HW a SW prostředí nutného pro zobrazení digitálních objektů. Bude vycházet z projektu TOTEM, který vytváří registr HW a SW prostředí pro jednotlivé formáty souborů, včetně přehledu nutných dynamických knihoven (DLL) pro jejich spuštění. Pro popis všech typů prostředí bylo v projektu vytvořeno metadatové schéma TOTEM, které v současnosti používají souběžně vedle PREMISu např. v Britské knihovně [KONSTANTELOS, 2012].

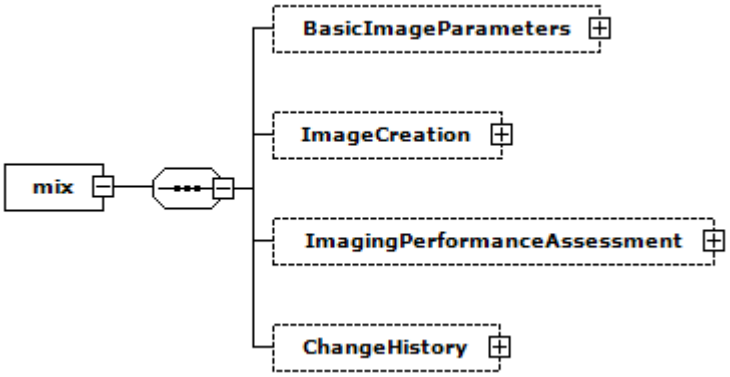
7.2.4 MIX

Vedle standardu PREMIS a jeho části Object bude pro popis technických metadat v návrhu profilu pro NDK použit také standard MIX. Ten je určen na popis technických vlastností obrazových souborů, ať vzniklých v digitalizaci, nebo digital-born. MIX je v současnosti k tomu účelu jedním z nejvíce používaných standardů. Záznam MIX lze vnořit do METS záznamu, s tímto přístupem se počítá v návrhu profilu metadat NDK. Od verze 2.0 lze MIX také vložit do záznamu PREMIS, pomocí elementu <extension>. Tímto způsobem MIX ukládá finská národní knihovna. MIX se často také používá jako zcela samostatný XML soubor, který se ukládá s ostatními metadaty. Tento přístup se používá v NK ČR od roku 2008.

Standard MIX v poslední verzi 2.0 má čtyři hlavní části (tedy elementy vrchní úrovně), které zobrazuje Obrázek 28 a jsou popsány níže:

- <basicImageParameters> – obsahuje základní technická metadata o digitálním objektu, např. formát, velikost, kontrolní součet apod.;
- <imageCreation> – obsahuje údaje o skenovacím přístroji nebo kameře, použitém SW, nastavení (clona apod.);

- <imagingPerformanceAssessment> – obsahuje informace nutné pro zobrazení objektu a kontrolu jeho kvality, např. rozměry, počty bitů, komprese, barevné nastavení atd.;
- <changeHistory> – uvádí všechny změny na objektu, na jakém SW a HW byly provedeny apod.



Obrázek 28 – Základní struktura MIX záznamu.

Vývoj standardu MIX se datuje až před rok 2000, standard samotný je v současnosti spojen s dokumentem *Data Dictionary – Technical Metadata for Digital Still Images*. Prvopočátky standardu se ale pojí ke specifikaci DIG35, která byla publikována v roce 2000 jako výstup pracovní skupiny *Digital Imaging Group*. Skupina chtěla vytvořit metadatový standard, který by dokázal popsat obrazový dokument z různých hledisek. Standard měl podporovat archivaci, indexaci, posuzování a vyhledávání jednotlivých obrazů. První verze DIG35 specifikace pod názvem *DIG35 Specification: Metadata for Digital Images, Version 1.0* [DIGITAL IMAGING GROUP, 2000] vyšla v září 2000. Dokument obsahoval úvod, příklady využití, přehled elementů a technické přílohy k implementaci a také XML DTD schéma. V roce 2001 vyšla nová verze 1.1 specifikace, která již byla postavená na XML (do té doby bylo XML pouze doporučené). Jedna z těchto verzí DIG35 byla použita v projektu *Manuscriptorium* v NK ČR pro nová metadata při přechodu na XML a standard MASTER+.

Ve stejné době vznikla aktivita americké organizace NISO (*National Information Standards Organization*) pod názvem *NISO Data Dictionary for Technical Metadata for Digital Still Images*. Draft dokumentu byl vydán ke komentářům v únoru 2001 a inspiroval se i specifikací DIG35. Úkolem bylo vytvořit slovník technických metadat, která by popsala vlastnosti obrazových digitálních objektů včetně procesu jejich vytvoření. Metadata tak měla podporovat hodnocení kvality, úpravy obrazů a také jejich dlouhodobou ochranu. Dosud byla vytvářena pouze metadata popisná pro vyhledávání. Cílem NISO iniciativy bylo vytvořit pro tuto mezeru datový slovník [COVER, 2002]. Součástí draftu NISO datového slovníku zatím nebylo žádné XML schéma, to bylo publikováno jako draft normy Z39.87 až v roce 2002 jako doprovodná část datového slovníku. Schéma dostalo název NISO MIX (*NISO Metadata for Images in XML*). První veřejná finální podoba MIX standardu je verze 0.2 z roku 2004. Poslední verze je 2.0 z roku 2010[285], která se od verze 0.2 dosti podstatně liší strukturou záznamů a pojmenováním elementů vrchní úrovně. Sémanticky jsou obě verze kompatibilní, strukturálně ovšem ne.

[285] http://www.loc.gov/standards/mix

První použití standardu MIX je v NK ČR spojeno s projektem *Manuscriptorium*, který začal od roku 2003 používat vlastní specifikaci technických metadat, postavenou částečně na specifikaci DIG35. V té době již existoval draft NISO datového slovníku, vůči kterému byla sada metadat doprovázející dokumenty v projektu *Memoriae Mundi Series Bohemica* analyzována. Další použití, tentokrát odpovídající přesně standardu MIX, se objevilo v roce 2007, kdy byla vytvořena specifikace administrativních metadat pro novodobé dokumenty. Ta obsahovala také specifikaci polí MIX – viz Tabulka 3 (s. 147). V ČR je standard MIX v jeho aktuální verzi používán také v Archivu hlavního města Prahy v projektu *Ad fontes*[286]. Záznam je ale velmi minimalistický (kontrolní součet, identifikátor, ICC profil, rozměry, barevný model, formát, údaje o skeneru, barevný profil, popis transformací a datum skenování) [HANOUSEK, 2010]. Dále se MIX používá na Univerzitě Karlově v SW pro repozitáře DigiTool.

V návrhu profilu pro metadata digitalizace v projektu NDK se počítá s využitím MIX standardu v rozsáhlé podobě, která je různá pro různé typy digitálních obrazů. Např. pro objekty ve formátu JPEG 2000 obsahuje návrh speciální elementy odlišné od těch, které bude mít např. popis digitálního objektu ve formátu JPEG. Rozsah elementů se může zdát příliš podrobný. Pokud by specifikace vznikala pouze pro procesy NK ČR a projektu NDK, lze si přestavit, že mohl být záznam MIX stručnější. Mnoho informací nám dodá na vstupu objektů do LTP systému sám systém, a to pomocí služeb třetích stran (JHOVE aj.). Specifikace je ovšem rozsáhlejší s ohledem na uložení metadat v knihovnách, které metadata nebudou posílat do LTP systému, ale pouze je spolu s jednotlivými obrazovými objekty uloží na svůj digitální repozitář, nebo pouze do úložiště s file systémem. Tyto knihovny pak budou mít dostupnou kompletní informaci, která bude využitelná i mimo LTP systém.
Technická metadata MIX musí vznikat automatickou cestou, manuálně by to působilo velké zdržení a spíše by to ani nešlo. Automatickou cestou se myslí pomocí nástrojů jako je JHOVE, Kakadu, NZME a jiných. Tyto nástroje vyčítají technická metadata z digitálních objektů a buď přímo vytvářejí záznam MIX, nebo je možné jejich výstupy do MIX převést. Některé údaje bude muset také automatickou cestou dodávat SW digitalizace a jednotlivých zařízení (skenerů, fotoaparátů)[287].
Nejpodstatnější pro LTP systém jsou údaje o vzniku a nastavení skeneru, které nástroje třetích stran, jako JHOVE apod., nejsou ze samotných digitálních objektů, zvláště po jejich úpravách, často schopny získat. V LTP systému by pak tento typ údajů chyběl, což zvláště u nastavení skenerů by byla škoda pro ilustraci celého životního cyklu konkrétních objektů.

7.2.5 ALTO XML

Standard ALTO XML má v návrhu profilu metadat pro projekt NDK odlišnou úlohu, než ostatní standardy. Ty mají za úkol popsat digitální dokument, ať z bibliografického nebo z technického, administrativního či strukturálního pohledu. ALTO XML (*Analyzed Layout and Text Object*) má za cíl uchovat metadata popisující vzhled a obsah fyzického textového dokumentu. Tím tedy pomoci zpřístupnění samotného digitálního objektu a to konkrétně naskenované stránky monografie

[286] http://www.ahmp.cz/index.html?mid=0&wstyle=0&page=adfontes/adfontes.html
[287] Ve výrobě novodobých dokumentů byly hodnoty pro jednotlivé elementy často napevno nastaveny přímo v SW, který vytvářel metadata. To pak působí problémy, když se tyto hodnoty neaktualizují, metadata poté neodpovídají realitě.

nebo periodika. Z tohoto pohledu je v metadatovém profilu pouze jakýmsi doplňkem, který nemusí být relevantní pro jiné typy digitálních objektů, které se někdy v budoucnu mohou dle tohoto metadatového profilu popisovat (např. audio aj.). ALTO XML má za úkol uživateli, který si prohlíží naskenovaný dokument, zvýraznit konkrétní oblasti (odstavce, slova) na zobrazené stránce tak, aby uživatel měl co největší komfort a viděl například vyhledávaný výraz v textu barevně označený. ALTO XML je tedy závislé na procesu rozeznávání znaků (OCR – *Optical Character Recognition*), který je jedním z kroků digitalizace. Výstup OCR procesu je čistý text, který je převeden do standardu ALTO XML, který rozeznaným výrazům přidá souřadnice a vytvoří tak jakousi mapu výrazů na konkrétní stránce. Díky tomu pak aplikace zpřístupnění může obě vrstvy, tedy obrázek naskenované strany a výrazy na ní v souřadnicích, zkombinovat. Vzhledem k tomu, že se jedná o rozeznávání textu, není ALTO XML relevantní pro dokumenty historické a z 19. století, které jsou psány českým nebo německým novogotickým písmem. V případě české mutace novogotického písma neexistuje nástroj, který by rozeznávání textů spolehlivě prováděl. Tyto dokumenty tedy v projektu NDK nebudou procházet procesem OCR, počítá se s ním pro budoucnost, až budou dostupné nástroje.

ALTO XML je v současné době ve správě Kongresové knihovny, stejně jako ostatní jmenované standardy. Do roku 2009 byla ovšem vývojem ALTO standardu komerční firma CCS GmbH, která spolu s dalšími institucemi (Univerzitní knihovna v Innsbrucku, národní knihovny Norska, Francie, Univerzitní knihovna Cornel, USA aj.) vyvinula standard v rámci evropského projektu METAe (*The Metadata Engine Project*; 2000-2003). ALTO XML se poté stalo součástí jejich komerčního SW na digitalizaci a workflow digitalizace DocWorks, který je využíván v projektech masové digitalizace v národních knihovnách Finska, Rakouska, Nizozemska a v mnohých univerzitních knihovnách USA. Základ SW DocWorks byl také výstupem projektu METAe a právě při vytváření tohoto SW se ukázalo, že neexistuje žádný standard pro záznam fyzické podoby skenované stránky (okraje, poloha znaků aj.). Aktuální verze ALTO XML je 2.0.
Každý ALTO XML záznam obsahuje sekci stylů (elementy <textStyle> a <paragraphStyle>), kde jsou vyjmenovány všechny styly (odstavce, fonty), které jsou na konkrétní stránce. ALTO XML záznam odpovídá vždy jedné naskenované stránce dokumentu. Sekce popisující podobu stránky obsahuje text stránky. Každá stránka je rozdělena na různé oblasti, jako jsou okraje a tiskové zrcadlo (Print space; Left margin; Right margin; Top margin; Bottom margin). Pro každou oblast jsou vyjmenovány objekty, které uvnitř ní byly rozeznány.
ALTO XML je navázáno, stejně jako předchozí standardy, na standard METS, přesněji na jeho strukturální mapu a v ní element <area>, který obsahuje odkazy do ALTO XML záznamu na konkrétní entity (články a v nich jednotlivé věty a v nich jednotlivá slova). Proto se někdy lze setkat s výrazem METS ALTO, který tuto symbiózu akcentuje. Odkaz může vypadat např. takto <area BETYPE="IDREF" FILEID="ALTO0005" BEGIN="P4_TB0002"/>. Soubor ALTO XML, na jehož část strukturální mapa odkazuje (jde o ALTO záznam stránky 4, textový blok číslo 2) musí být samozřejmě odkazován jako celek v METS části <fileSec>. Část vlastního ALTO XML záznamu, na kterou METS odkazuje, může vypadat jako Ukázka 14. Ukázka popisuje souřadnice textového bloku 2 (TB – *Text Block*) na stránce 4, souřadnice řádky textového bloku (*textline*) a také souřadnice prvního slova řádky (*string*). První slovo je „Jaro". Tento přístup je i v profilu pro NDK. ALTO XML lze použít i jako samostatný soubor bez METS záznamu. Z uvedeného vyplývá, že v projektu NDK bude vznikat ALTO na úroveň slova. Je možné vytvářet ALTO XML i na úroveň

znaku, je to ale časově náročnější, zvyšuje to velikost XML souboru a v případě NDK taková granularita není potřeba.

```
<TextBlock ID="P4_TB00002" HPOS="230" VPOS="807" WIDTH="1371" HEIGHT="2816" STYLEREFS="TXT_1 PAR_BLOCK">
  <TextLine ID="P4_TL00002" HPOS="330" VPOS="810" WIDTH="1257" HEIGHT="47">
    <String ID="P4_ST00006" HPOS="330" VPOS="810" WIDTH="85" HEIGHT="39" CONTENT="Jaro"/>
```

Ukázka 14 – Část ALTO XML záznamu textového bloku, řádky a slova.

7.3 Poznámky ke konkrétním aspektům navrhovaného aplikačního profilu metadat pro NDK

7.3.1 Uživatelské kopie a jejich metadatový popis

Specifikace metadatového profilu pro NDK popisuje pouze archivní kopie. Uživatelské kopie metadata nepopisují, ale jejich soubory jsou součástí balíčku vzniklého v digitalizaci. Důvodem rozhodnutí je skutečnost, že již dnes víme, že uživatelské kopie budou v budoucnu nahrazeny novými uživatelskými kopiemi. K tomu dojde v okamžiku, kdy nastane změna formátu a současné uživatelské kopie nebude možné uživatelům nabízet. Není tedy nutné věnovat uživatelskými kopiím stejnou pozornost jako kopiím archivním. Není třeba vytvářet podrobná technická metadata, nebudeme je totiž potřebovat k plánování ochrany ani k ochranným akcím. Uživatelské kopie v NDK budou ukládány mimo LTP systém, a to přímo na serverech aplikací zpřístupnění (Kramerius, Manuscriptorium). Tyto aplikace mohou využít ostatních metadat METS záznamu archivních kopií, např. strukturálních a popisných tak, že linky na digitální objekty v záznamu, které obsahují archivní kopie, nahradí za linky na kopie uživatelské. To mnohdy znamená pouze drobnou úpravu názvu souboru (tedy např. MC_123456.jp2 se nahradí za UC_123456.jp2). V době přípravy profilu existovala i varianta s vlastním METS záznamem pro uživatelské kopie, byla ale opuštěna z důvodů uvedených výše.[288] Argument, že tvorba uživatelských kopií je nákladná věc, je jistě pravdivý, ovšem je velmi pravděpodobné, že nová verze uživatelských kopií bude vznikat z kopií archivních, ne z předchozích kopií uživatelských. Ochrana bitstreamu uživatelských kopií je zajištěna dostatečným zálohováním aplikací zpřístupnění. Pro zvolený přístup hraje i skutečnost, že i kdybychom se později rozhodli uživatelské kopie do LTP systému vkládat, tak rozsáhlá technická metadata vzniknou automatickými procesy během jejich vkládání (modul Ingest).

Při vstupu do LTP systému se ani technická metadata archivních kopií, jak je specifikuje metadatový profil NDK, nevyužijí v podobě, v jaké jsou. Na vstupu vzniknou nová, ovšem ta původní mohou sloužit ke kontrole oproti nově vytvořeným. Velmi důležitá jsou metadata administrativní, o událostech, které nelze vytvořit automaticky. To jsme si vyzkoušeli v POC v roce 2010, kdy z DTD se technická metadata nepoužila, ale údaje o vzniku objektů chyběly. Důvodem proč specifikace má velké množství technických metadat je snaha poskytnout co nejvíce údajů pro instituce, které nebudou používat LTP systém, ale třeba pouze file systém. Pak mohou s daty pracovat, dělat analýzy. V případě pozdějšího uložení do LTP mohou dělat kontroly na vstupu.

[288] Dalším důvodem bylo i to, že aplikace zpřístupnění není na METS záznam připravena, nemělo by smysl jej tedy vytvářet v digitalizaci, dále by se totiž nevyužil.

7.3.2 Duplikace elementů různých schémat

Pokud v jednom záznamu METS zkombinujeme několik schémat metadat, je pravděpodobné, že se určité elementy mohou opakovat a překrývat. Je to případ standardu PREMIS (část Object), pokud popisuje obrazový dokument, a standardu pro technická metadata digitálních obrazů MIX, kde se mohou opakovat např. kontrolní součty, údaje o formátu objektu apod. Oba standardy jsou v různých <techMD> částech METS záznamu, která se pro každé schéma opakuje, a to s různými atributy (MDTYPE apod.) k jejich odlišení. Redundantní výskyt může nastat také v případě, že stejný element je definován v METS (často jako atribut) a v dalším schématu, který je do něj vložen. Tato situace nastává s PREMISem. Kupříkladu METS i PREMIS obsahují údaje o formátu dat, ale každý za jiným účelem. PREMIS umožňuje linkování na PRONOM, zatímco METS ne. Podobně se může opakovat údaj o velikosti souboru, kontrolním součtu – více viz kapitola 7.2.1.3.
Je potřeba udělat rozhodnutí, zda tyto duplikace budeme akceptovat a mít je v obou schématech, nebo pouze v jediném z nich. Pokud se rozhodneme nechat konkrétní údaje duplikované, není to na škodu, je pouze nutné vytvořit pravidla, která zajistí, že procesy strojového zpracování (tedy konkrétní SW aplikace) si toho budou vědomy a budou vědět, kterému výskytu shodného údaje dát přednost.
Míchání schémat je v dnešní době, i díky METS, běžné a cílené. Není již důvod vyvíjet vlastní schéma obsahující všechny potřebné elementy. Kdysi se schémata takto vytvářela, a pak bylo samozřejmě možné dosáhnout toho, že se elementy ani hodnoty v nich neopakovaly. Dnešní přístup je jiný, ať se hodnoty klidně opakují, jen pokud jsou součástí různých schémat vložených v METS záznamu. METS záznam je pouze drží pohromadě, ale využití je plánováno pro každé schéma zvlášť. Výhodou přístupu násobného výskytu je, že lze záznamy jednotlivých standardů dále použít i nezávisle na METS záznamu, aniž by jim chyběly podstatné údaje kvůli tomu, že se již objevily v jiném schématu, který byl součástí původního společného METS záznamu. Do úvahy je třeba vzít např. i to, že záznam METS je často určen pro zobrazení, kdežto např. PREMIS naopak pro ochranu, měl by tedy obsahovat maximum údajů.
Validace jednotlivých schémat v METS záznamu nepředstavuje problém. METS je XML schéma uzpůsobené na vložení jednoho nebo více dalších XML záznamů dalších schémat. Aby se odlišily jednotlivé kontexty elementů se stejným jménem (např. autor), jsou součástí METS záznamu deklarace jmenných prostorů XML (xmlns) jako atributy kořenového elementu. Každý jmenný prostor je identifikován pomocí URI. Na toto téma viz např. [ZENG a QIN, 2008, s. 137]. Pokud by jmenné prostory nebyly definovány, obsah shodných polí nebude možné sémanticky odlišit a jednotlivé systémy si s takto vzniklým nepořádkem nebudou schopny poradit. Dojde ke kolizím systému např. při výměně dat s jiným systémem.

7.3.3 Struktura balíčku a záznamu metadat

Specifikace profilu pro NDK digitalizaci je záměrně udělána tak, že všechny typy metadat balí do standardu METS. Důvodem je to, že pokud podle specifikace bude postupovat knihovna (např. v rámci VISK7), která nemá LTP systém a bude ukládat výsledky digitalizace do file systému, je pro ni lepší, když má metadata logicky zabalená pomocí METS. Metadata tak mají strukturu, jsou ve standardním formátu a dají se bez problémů skladovat, protože METS kontejner je drží logicky pohromadě. Tento postup je využíván velmi často.

Další možností by byl postup, kdy by popisná, technická, strukturální a administrativní metadata byla ve svých nativních formátech (MODS, PREMIS, MIX aj.) uložena ve složce jako jednotlivé XML záznamy. Podle pojmenování XML souborů by bylo jasné, ke kterému obrázku patří a o jaký typ metadat jde (např. MIX_123.xml by označoval MIX metadata k obrázku číslo 123). Tento přístup není vhodný k dlouhodobému uložení, nedrží metadata příliš pohromadě. Nevhodný je také pro přemísťování a manipulaci s metadaty, protože je nutno manipulovat několik XML záznamů namísto jednoho METS záznamu. To jsme si v NK ČR vyzkoušeli na minulých generacích digitalizovaných dokumentů, které byly uloženy přesně takto.
Obecně struktura balíčku přicházejícího z digitalizace není pro LTP systém příliš důležitá. Podstatné je, zda obsahuje všechna potřebná metadata. Zda jsou v METS nebo v jednotlivých XML záznamech je LTP systému jedno. Vždy bude potřeba provést mapování do jeho vnitřního formátu.

Hlavní odchylkou od očekávané struktury METS záznamu čísla periodika nebo svazku monografie je skutečnost, že administrativní metadata nejsou součástí hlavního METS záznamu, ale jsou ve vedlejším METS záznamu zvlášť. Důvodem je zkrácení hlavního METS záznamu, dále možnost rychlé manipulace a využití administrativních metadat konkrétních naskenovaných obrazů v podobě vedlejšího METS záznamu bez nutnosti manipulovat s hlavním METS záznamem. Stejné řešení je použito národní knihovně Norska (PREMIS a MIX mimo hlavní METS záznam) a částečně i v národní knihovně Nizozemí (PREMIS je mimo hlavní METS záznam). Právě z příkladů těchto digitalizací obě specifikace profilu vycházejí.
Vedlejší METS záznamy jsou uvedeny v části <fileSec> hlavního METS záznamu. METS standard umožňuje, že soubory uvedené v části <fileSec>, tedy obrázky, OCR soubory aj. mohou mít vazbu na administrativní metadata v části <amdSec>. Vazba je vyjádřena pomocí atributu ADMID. V případě profilu pro NDK ovšem v hlavním METS záznamu čísla nebo svazku část <amdSec> chybí. Není to problém, protože tato část není povinná. Administrativní a technická metadata jsou ve vedlejším METS záznamu. Hlavní METS ví o vedlejším METS záznamu (XML souboru) díky tomu, že je uveden v jeho části <fileSec> jako další soubor, který tvoří digitální objekt, stejně jako obrazy a OCR soubory. Vazbou mezi konkrétním souborem a jeho administrativními metadaty je název vedlejšího METS záznamu. Např. AMD_METS_ANL_123456_0013.xml označuje stránku 13 konkrétního čísla periodika s identifikátorem 123456. Název relevantního souboru k těmto metadatům je např. ANL_123456_0013.jp2 pro archivní kopii, ANL_123456_0013.txt pro textový soubor OCR apod. Vedlejší záznam METS je tak spojen se všemi reprezentacemi stránky 13, která je uvedena v hlavním METS záznamu v části <fileSec>. Vedlejší METS je linkován z <fileSec>, protože specifikace pro NDK všechny vložené záznamy metadat v jiných schématech do METS záznamu balí, neodkazuje je. Nebylo možné odkázat na vedlejší METS např. z hlavního METS záznamu a jeho části <amdSec>. Odkaz totiž musí vést pouze na XML záznam ve schématu PREMIS, MIX, ovšem ne na záznam METS. Odkaz přes <fileSec> navíc jasně ukazuje, že vedlejší METS je součástí celého balíčku dat a metadat.

8. Závěr

Paměťové instituce, především knihovny a archivy, se v digitálním světě rychle proměňují. Informace, které do nich jsou ukládány, již dávno nejsou pouze na fyzických nosičích, ale přicházejí stále více v digitální podobě. V současnosti již není možné tento trend ignorovat. Světová komunita knihoven a archivů aktuálně řeší, jak se s digitálními dokumenty vypořádat a především jak je zachovat pro budoucnost, podobně jako zajišťovala uchování fyzických dokumentů.

Jedním ze zdrojů digitálních dokumentů je digitalizace sbírek, v rámci níž knihovny i archivy převádějí své analogové dokumenty do digitální podoby. Rozvoj digitalizace začal pomalu v druhé polovině 90. let předcházejícího století. Po roce 2000 projektů digitalizace přibývalo, zlepšovaly se technologie, ustálily se standardy. Zvyšovalo se logicky také množství dokumentů, které byly zdigitalizovány a byly ukládány v digitálních repozitářích. Po roce 2005 přišla ke slovu tzv. masová digitalizace za využití robotických skenerů, čímž se rapidně zvýšila produkce digitalizovaných dokumentů/objektů/stran. Jeden robotický skener je totiž schopný naskenovat až 2000 stran za hodinu. Od roku 2005 bylo již většině institucí jasné, že digitalizace je pouhým začátkem celého procesu. Může řešit ochranu fyzických předloh, ale vytvoření digitální kopie znamená také řešit ochranu a uchování digitálního dokumentu, což je nesrovnatelně komplikovanější, než u dokumentu fyzického.

Instituce začaly řešit problematiku dlouhodobé ochrany digitálních dat, a to nejen ve smyslu ochrany bitstreamu pomocí záloh a přesunu na stále nová a nová média. Podstatou se ukázala být logická dlouhodobá ochrana digitálních dat, která zajišťuje použitelnost digitálních objektů v budoucnu. Použitelnost spočívá v tom, že digitální objekt lze vyhledat, zobrazit, a uživatel je schopen jej pochopit. Má k dispozici kontextové údaje (metadata), které mu řeknou, o jaký objekt jde, kdy a proč vznikl a zda prošel za dobu uložení nějakými změnami. Pokud změnami prošel, musí být jasné jakými a zda při změnách ztratil některé ze svých vlastností. Uživatel si musí být jistý, že s digitálním objektem nebylo záměrně manipulováno, nebyl měněn jeho obsah ani podoba. Budoucí uživatel musí mít v digitální repozitář a instituci, která jej provozuje, maximální důvěru. Metadata nejsou jedinou, i když významnou, zárukou důvěryhodnosti. Důvěryhodnost podtrhuje fungování celého digitálního repozitáře a instituce, která je schopná prokázat, že procesy, které s digitálními objekty provádí, jsou standardní a dokumentované.

Z textu kapitol 3 a 4 jasně vyplývá, že pro logickou dlouhodobou ochranu digitálních dat jsou klíčová hlavně metadata technická, administrativní a ochranná. Důležitost tvorby těchto typů metadat se jako nosné téma prolíná celou disertační prací. Popis vývoje použití metadat v NK ČR v kapitole 5 je do jisté míry dokreslením úvodní části o metadatech. Na pozadí vývoje používání metadat byl ilustrován jejich smysl v projektech digitalizace i to, jak se tento smysl v průběhu doby měnil. V počátcích šlo vždy o digitalizaci „pro ochranu“, později o digitalizaci „pro zpřístupnění“. V obou případech vždy metadata která vznikala, měla pomoci k popisu a hlavně vyhledání konkrétního digitálního objektu. V tomto konceptu nastala po roce 2005 výrazná změna. Začaly se objevovat nové typy metadat, jejichž smysl byl odlišný od popisných schémat. Cílem nebylo dokumenty popsat z pohledu bibliografického, ale popsat je z pohledu ochrany, technických vlastností a také z pohledu změn a procesů, které se s digitálními dokumenty během jejich životního cyklu děly. Tato změna byla reflektována v celé komunitě, začaly vznikat nové metadatové standardy (PREMIS, MIX, LMER a mnoho dalších). Proměnil se i proces tvorby metadat v digitalizaci, kdy začalo být nutné tyto nové typy metadat vytvářet. Cílem bylo zajištění předpokladu pro následné procesy dlouhodobé ochrany v digitálních repozitářích. Začaly vznikat

specializované systémy na uložení digitálních dat, které dále ochranná, technická a administrativní metadata doplňovaly, a to při jakékoliv změně tak, aby bylo jasné, co se během životního cyklu s dokumentem dělo. Kapitola 4 popisující vývoj a typy metadat ukazuje, že možností přístupů k metadatům nebo výběru konkrétních standardů je mnoho a může být těžké se v nich zorientovat.

V prostředí českých knihoven a archivů nebyla potřeba tvorby nových typů metadat zcela jednoznačně reflektována a to ani v probíhajících projektech, jak ukazují komentáře v kapitole 5. V NK ČR vznikla první specifikace technických a ochranných metadata až v roce 2008 a to ve velmi základní podobě. Dalo by se mluvit spíše o pilotním projektu tvorby těchto metadat. Tento rozsah vydržel až do současnosti, kdy od zmíněného roku v rámci projektu VISK7 vznikají základní metadata ve standardech PREMIS a MIX u digitalizovaných novodobých děl. Díky nástroji na digitalizaci Sirius, byla základní sada ochranných, administrativních a technických metadat produkována i v dalších projektech, které tento SW používaly nebo používají (např. Městská knihovna v Praze). Od roku 2008 ovšem nevznikla v oblasti českého knihovnictví a digitalizace žádná další iniciativa, která by tento typ metadat dále rozpracovala. Je to dáno i tím, že v ČR neexistoval a dosud neexistuje repozitář, který by tento typ metadat dokázal využít a následně dále vytvářet, jak vyplývá z kapitoly 6. Bez ohledu na tuto skutečnost jsou ochranná metadata velmi důležitá, i v případě, že není dostupný systém, který by je využil. Pokud jsou digitální objekty dobře technicky i jinak popsány, je vždy možné lépe kontrolovat jejich integritu, autenticitu. Bez těchto metadat nelze jednoznačně říci, zda digitální objekt neprošel nechtěnou změnou, není poškozen, není odlišný od původního objektu apod. Bohužel se zdá, že v ČR dosud tento typ problémů nenašel větší odezvu a tvorba ochranných metadat v digitalizaci stále není prioritou.

Jistým obratem k lepšímu je až projekt Národní digitální knihovna, v rámci kterého vznikl návrh nových metadatových profilů pro digitalizaci monografií a periodik. Návrhy, které jsou praktickou částí této disertační práce, vznikaly intenzivně během let 2009-2011. Obsahují všechny potřebné výše uvedené typy metadat, které lze převážně automatizovaným způsobem tvořit v procesu digitalizace. Jde o návrhy záměrně flexibilní a postavené pouze na schématech metadat, která jsou ve své oblasti ve světě běžně používanými standardy. Díky tomu je možné metadata sdílet, vytvářet a upravovat v mnoha běžně dostupných nástrojích, které s konkrétními standardy pracují apod. Již dnes je jasné, že v blízké budoucnosti bude potřeba obdobný návrh profilu metadat vytvořit i pro jiné typy digitálních dokumentů, např. pro digital-born dokumenty, elektronický povinný výtisk (viz projekt NK ČR projekt pro správu elektronických publikací *eDeposit*), data vzniklá archivací webu aj. Návrh profilu NDK by pro tyto snahy měl být dobrým východiskem i vodítkem.

Věřím tomu, že předložená disertační práce odpovídá na otázky a plní cíle, které stály na počátku jejího vzniku. Návrh aplikačního profilu metadat je přínosem pro oblast digitalizace a tvorby metadat v České republice a doufám, že bude dále rozvíjen a využíván nejen v projektu NDK[289], ale i v projektech jiných, jak tomu ostatně je v projektu digitalizace článků periodiky ANL+.

[289] 21. 3. 2012 byly publikovány nové, mírně aktualizované, verze obou profilů. Na této úpravě se původní autor (Jan Hutař) již z větší části nepodílel – viz http://kramerius-info.nkp.cz/planovane-akce/novinky/zverejneny-aktualizovane-verze-metadatovych-formatu).

9. Seznam použitých zdrojů a literatury

Seznam použité literatury je řazen abecedně dle prvního údaje v záznamu (záhlaví), seznam není číslován. Pro lepší přehlednost je v seznamu literatury každé záhlaví zvýrazněno. Jednotlivé záznamy jsou v souladu s pravidly uvedenými v normě ČSN ISO 690 (01 0197) platné od 1. dubna 2011. Výjimkou je uvádění roku vydání, který je uveden oproti normě dvakrát, jednou za záhlavím a jednou za údaji o vydavateli.

ABBOTT, Daisy. 2003. Overcoming the Dangers of Technological Obsolescence: rescuing the BBC Domesday Project. *DigiCULT.info: a newsletter on Digital Culture* [online]. August 2003, issue 4, s. 7-10 [cit. 2011-05-25]. ISSN 1609-3941. Dostupné z: http://www.digicult.info/downloads/digicult_newsletter_issue4_highres.pdf

ABID, Abdelaziz. 2007. Safeguarding our digital heritage: a new preservation paradigm. In: LUSENET, Yola de a Vincent WINTERMANS (eds.). *Preserving the digital heritage: principles and policies*. Amsterdam: Netherlands National Commission for UNESCO, European Commission on Preservation and Access, 2007, s. 7-14. ISBN 978-90-6984-523-4. Dostupné také z: http://www.ica.org/5697/paag-resources/preserving-the-digital-heritage-principles-and-policies.html

AIP Beroun. 2005. *Manuscriptorium v. 2.0: komplexní digitální dokument.* Draft, verze 1.0. Praha: Národní knihovna ČR, 2005. 55 s. Dostupné také z: http://digit.nkp.cz/projekty/VZ-2004_2010/2005/ManuscriptoriumKDD.pdf

ALTENHÖNER, Reinhard. 2009. *Osobní rozhovor při návštěvě Národní knihovny Německa*. Frankfurt nad Mohanem, 23. duben 2009.

ARTHUR, Kathleen, et al. 2004. *Recognizing Digitization as a Preservation Reformatting Method.* Chicago: Association of Research Libraries, 2004. 17 s. Dostupné z: http://www.arl.org/bm~doc/digi_preserv.pdf

ASSOCIATION FOR LIBRARY COLLECTIONS & TECHNICAL SERVICES. 2000. Task Force on Metadata: Summary Report 1999. In: *American Library Association* [online]. American Library Association, 27. 8. 2000 [cit. 2011-12-14]. Dostupné z: http://www.ala.org/cfapps/archive.cfm?path=alcts/organization/ccs/ccda/tf-meta3.html

ASSOCIATION FOR LIBRARY COLLECTIONS & TECHNICAL SERVICES. 2007. Definitions of Digital Preservation. In: *American Library Association* [online]. Americal Library Association, 2007 [cit. 2011-04-14]. Dostupné z: http://www.ala.org/ala/mgrps/divs/alcts/resources/preserv/defdigpres0408

BACA, Murtha. 2008. *Introduction to Metadata.* 2nd Edition. Los Angeles (CA): Getty Publications, 2008. 96 s. ISBN 978-0-89236-896-9. Dostupné také z: http://www.getty.edu/research/publications/electronic_publications/intrometadata/index.html

BAUEROVÁ, Zuzana. 2006a. MinervaPlus a MichaelPlus. In: *Archivy, knihovny, muzea v digitálním světě 2005.* Praha: Národní technické muzeum Praha, 2006, s. 13-14. Rozpravy Národního technického muzea v Praze, 195. ISBN 80-7037-149-8. Dostupné také z: http://skip.nkp.cz/KeStazeni/Archivy05/Sbornik_2005.pdf

BAUEROVÁ, Zuzana. 2006b. Plány Evropské komise - Evropská digitální knihovna. In: *Archivy, knihovny a muzea v digitálním světě 2005.* Praha: Národní technické muzeum, 2006, s. 15-17. Rozpravy Národního technického muzea v Praze, 195. ISBN 80-7037-149-8. Dostupné také z: http://skip.nkp.cz/KeStazeni/Archivy05/Sbornik_2005.pdf

BEARMAN, David. 1999. Reality and Chimeras in the Preservation of Electronic Records. *D-Lib Magazine* [online]. April 1999, vol. 5, no. 4 [cit. 2011-09-11]. ISSN 1082-9873. DOI 10.1045/april99-bearman. Dostupné z: http://www.dlib.org/dlib/april99/bearman/04bearman.html

BERNERS-LEE, Tim. 1997. Web architecture: Metadata. In: *W3C* [online]. 6. January 1997 [cit. 2011-10-14]. Dostupné z http://www.w3.org/DesignIssues/Metadata.html

BONIN, Sonja. 2009. *Preservation and Long-term Access via NETworked Services: keeping digital information alive for the future* [online]. PLANETS Consortium, 2009 [cit. 2011-12-04]. 20 s. Dostupné z: http://www.planets-project.eu/docs/comms/PLANETS_BROCHURE.pdf

BORBINHA, José (ed.), et al. 2011. *iPRES 2011: 8th International Conference on Preservation of Digital Objects, 1.-4.11.2011, Singapore*. Singapore: National Library Board Singapore & Nanyang Technology University, 2011. 287 s. ISBN 978-981-07-0441-4. Dostupné také z: http://getfile3.posterous.com/getfile/files.posterous.com/temp-2012-01-02/dHqmzjcCGoexvmiBzJDCyhrlhIgswoffzvsfnpEAxjHFEesarvwahEHrmyvj/iPRES2011.proceedings.pdf

BRADLEY, Kevin. 2006. *Risks Associated with the Use of Recordable CDs and DVDs as Reliable Storage Media in Archival Collections - Strategies and Alternatives.* Paris: UNESCO, 2006. 28 s. Dostupné z: http://unesdoc.unesco.org/images/0014/001477/147782E.pdf

BRATKOVÁ, Eva. 1999. Metadata jako nový nástroj pro komunikaci webovských informačních zdrojů. *Národní knihovna: knihovnická revue*. 1999, roč. 10, č. 4, s. 178-195. ISSN 1214-0678. Dostupné také z: http://knihovna.nkp.cz/pdf/9904/9904178.pdf

BRITISH LIBRARY. 2008. *The British Library's Strategy 2008–2011: summary* [online]. London: British Library, 2008 [cit. 2011-12-13]. 2 s. Dostupné z: http://www.bl.uk/aboutus/stratpolprog/strategy0811/strategysummary.pdf

BRITISH LIBRARY. NATIONAL PRESERVATION OFFICE. 1997. *Preservation and Digitisation: principles, practice and policies: Papers given at National Preservation Office 1996 Annual Conference, University of York, 3-5 September.* London: National Preservation Office, British

Library, 1997. 101 s. ISBN 0-7123-4581-7. Dostupné také z: http://www.bl.uk/blpac/pdf/conf1996.pdf

BÜLOW, Anna a Jess AHMON. 2011. *Preparing Collections for Digitization.* London: Facet, National Archives, 2011. xvii, 184 s. ISBN 978-1-85604-711-1.

BURNARD, Lou. 2008. *ENRICH Schema: a Reference Guide* [online]. October 2008 [cit. 2011-11-01]. vi, 306 s. Dostupné z: http://projects.oucs.ox.ac.uk/ENRICH/Deliverables/referenceManual_en.pdf

BURNARD, Lou, James CUMMINGS a Sebastian RAHTZ. 2008. *Revised TEI conformant Specification* [online]. August 2008 [cit. 2011-03-01]. 137 s. Dostupné z: http://projects.oucs.ox.ac.uk/ENRICH/Deliverables/ENRICH_D_3_1_TEI-spec.pdf

BYERS, Fred R. 2003. *Care and Handling of CDs and DVDs: A Guide for Librarians and Archivists.* Washington: Council on Library and Information Resources, National Institute of Standards and Technology, 2003. 42 s. ISBN 1-932326-04-9. Dostupné také z: http://www.clir.org/pubs/reports/pub121/pub121.pdf

CANDELA, L., et al. [2006]. *The Digital Library Manifesto*. DELOS, [2006]. 20 s. ISSN 1818-8044. ISBN 2-912335-24-8. Dostupné také z: http://146.48.87.21/OLP/UI/1.0/Disseminate/1332139174V0c4em3JWM/a221332139174DYf42SQK

CANTARA, Linda. 2005. METS: the Metadata Encoding and Transmission Standard. *Cataloging & Classification Quaterly.* 2005, Vol. 40, Issue 3-4, Special Issue: Metadata: A Cataloger's Primer, s. 237-253. ISBN 978-0-7890-2801-3.

CAPLAN, Priscilla. 2003. *Metadata Fundamentals for All Librarians.* Chicago: American Library Association, 2003. ix, 192 s. ISBN 0-8389-0847-0.

CAPLAN, Priscilla, Angela DAPPERT a Markus ENDERS. 2010. *PREMIS Tutorial: understanding & Implementing the PREMIS Data Dictionary for Preservation Metadata: PREMIS Tutorial, iPRES 2010 Vienna, Austria, September 19, 2010* [online prezentace]. Vienna, 19. September 2010 [cit. 2011-11-13]. 33 snímků (ve formátu PPT). Dostupné z: http://www.loc.gov/standards/premis/premis-Vienna-pt1.pdf

CAPLAN, Priscilla a Carol C.H. CHOU. 2011. DAITSS Grows Up: Migrating to a Second Generation Preservation System. In: *Archiving 2011, May 16-19, 2011, Salt Lake City Utah: Final Program and Proceedings.* Springfield (VA): Society for Imaging Science and Technology, 2011, s. 101-104. ISBN 978-0-89208-294-0.

CAPLAN, Priscilla a Kate ZWAARD. 2011. Implementing PREMIS to support digital preservation. In: *Archiving 2011, 16-19.5.2011, Salt Lake City* [workshop]. Salt Lake City (UT), 16. květen 2011. Autor se workshopu účastnil.

CARPENTER, Leona. 2005. *Repositories in Context: Digital Repositories as components of an integrated infrastructure for education* [online prezentace]. 2005 [cit 2011-04-15]. Dostupné z: http://www.ukoln.ac.uk/events/delos-rep-workshop/presentations/carpenter.ppt

CELBOVÁ, Ludmila. 2003. DTD. In: *KTD: Česká terminologická databáze knihovnictví a informační vědy (TDKIV)* [online]. Praha: Národní knihovna ČR, 2003 [cit. 2011-04-02]. Sysno 000000526. Dostupné z: http://aleph.nkp.cz/F/?func=direct&doc_number=000000526&local_base=KTD

CENTER FOR RESEARCH LIBRARIES. 2007. Ten Principles. In: *Center for Research Libraries* [online]. 2007 [cit. 2011-04-16]. Dostupné z: http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying/core-re

COLE, Timothy W. 2002. Creating a Framework of Guidance for Building Good Digital Collections. *First Monday* [online]. May 2002, Vol. 7, Nr. 5 [cit. 2011-12-29]. Dostupné z: http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/rt/printerFriendly/955/876

Conference Day 1 & 2: a Brief Conference Report. In: Van der LAAN, M. J. *Strategies for a European Area of Digital Cultural Resources: towards a continuum of digital heritage: European conference, The Hague, The Netherlands 15-16 September 2004.* Haag: [s.n.], 2004, s. 6-13.

CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEMS. 2002. *Reference Model for an Open Archival Information System (OAIS): CCSDS 650.0-B-1* [online]. Washington (DC): Consultative Committee for Space Data Systems, January 2002 [cit. 2011-11-05]. 148 s. Dostupné z: http://public.ccsds.org/publications/archive/650x0b1.PDF

COVER, Robin. 2002. DIG35: Metadata Standard for Digital Images. In: *Cover Pages* [online]. 10. červen 2002 [cit. 2011-07-22]. Dostupné z: http://xml.coverpages.org/dig35.html

CROSSCZECH. [2010]. *Národní digitální archiv: studie proveditelnosti*. Praha: Národní archiv České republiky, [2010] [cit. 2011-11-27]. 183 s. Dostupné také z: https://web.nacr.cz/zakazky/NDA_projekt_ISNDA/dokumenty/NDA_IS_ZD_priloha1.zip

CUBR, Ladislav, Jan HUTAŘ a Marek MELICHAR. 2008. Stav implementace perzistentních identifikátorů v NK ČR a výhled do budoucnosti. In: *1. ročník semináře zaměřeného na problematiku uchovávání a zpřístupňování šedé literatury, Praha 8. 10. 2008* [online]. Praha: Národní technická knihovna, 2008 [cit. 2011-07-12]. 4 s. ISSN 1803-6015. Dostupné z: http://nusl.techlib.cz/images/PID_text.pdf

CUNDIFF, Morgan a Nate TRAIL. 2007. Using <METS> and <MODS> to Create XML Standards-based Digital Library Applications. In: *American Library Association Congress* [online]. Washington (DC): Library of Congress, 2007 [cit. 2011-10-08]. 42 snímků (ve formátu PPT). Dostupné z: http://www.powershow.com/view/2eb59-NzkwY/Using_METS_and_MODS_to_Create_an_XML_Standardsbased_Digital_Library_Application_flash_ppt_presentation

CUTTER, Charles Ammi. 1889. *Rules for a dictionary catalog.* Special Report on Public Libraries - Part II. Washington: Government Printing Office, 1889. 133 s. Dostupné z: http://www.archive.org/details/cu31924029519026.

ČESKO. 2004. *Státní informační a komunikační politika: e-Česko 2006* [online]. Praha, 2004 [cit. 2011-05-13]. 35 s. Dostupné z: http://knihovnam.nkp.cz/docs/SIKP_def.pdf

ČESKO. MINISTERSTVO FINANCÍ. 2011. Výzva č. 1 - SCHVÁLENÉ PROJEKTY 2007/2006. In: *Ministerstvo financí ČR* [online]. 11. červen 2011 [cit. 2011-08-07]. Dostupné z: http://www.mfcr.cz/cps/rde/xchg/mfcr/xsl/fm_norska_28797.html

ČESKO. MINISTERSTVO KULTURY. 2004. *Koncepce rozvoje knihoven v České republice na léta 2004 až 2010* [online]. Praha: Ministerstvo kultury České republiky, 2004 [cit. 2011-05-03]. 43 s. Dostupné z: http://knihovnam.nkp.cz/docs/Koncepce04_10.doc

ČESKO. MINISTERSTVO KULTURY. 2005. *VISK: Cíle programu* [online]. 11. leden 2005 [cit. 2011-05-23]. Dostupné z http://visk.nkp.cz/VISKcile.htm#cil

ČESKO. MINISTERSTVO VNITRA. 2009. Vyhláška o podrobnostech výkonu spisové služby. In: *Sbírka zákonů České republiky.* 2009, částka 57, s. 2773 - 2782. Dostupné také z: http://aplikace.mvcr.cz/sbirka-zakonu/ViewFile.aspx?type=c&id=5503

ČESKO. MINISTERSTVO VNITRA. ODBOR ARCHIVNÍ SPRÁVY. 2011. Archivní standardy - Ministerstvo vnitra České republiky. In: *Ministerstvo vnitra České republiky* [online]. Praha: Ministerstvo vnitra České republiky, 2010 [cit. 2011-09-18]. Dostupné z: http://www.mvcr.cz/clanek/archivni-standardy.aspx.

DAPPERT, Angela. 2009. *Digital Preservation Metadata* [online prezentace]. WePreserve, March 2009 [cit. 2011-10-20]. 122 snímků (ve formátu PPT). Dostupné z: http://www.slideshare.net/DigitalPreservationEurope/preservation-metadata-1258027

DIGITAL CURATION CENTRE. 2010. DCC Curation Lifecycle Model. In: *Digital Curation Centre* [online]. Edinburgh: Digital Curation Centre, 2010 [cit. 2011-12-30]. Dostupné z: http://www.dcc.ac.uk/resources/curation-lifecycle-model

DIGITAL IMAGING GROUP. 2000. *DIG35 Specification: metadata for Digital Images. Version 1.0* [online]. Digital Imaging Group, 2000 [cit. 2011-05-10]. 165 s. Dostupné z: http://xml.coverpages.org/FU-Berlin-DIG35-v10-Sept00.pdf

DIGITAL LIBRARY FEDERATION. 2007. *<METS> METADATA ENCODING AND TRANSMISSION STANDARD: PRIMER AND REFERENCE MANUAL* [online]. Digital Library Federation, 2007 [cit. 2011-07-19]. Dostupné z: http://www.loc.gov/standards/mets/METS%20Documentation%20final%20070930%20msw.pdf

DIGITAL PRESERVATION COALITION. 2009. Introduction: Definitions and Concepts. In: *DPC Digital Preservation Handbook* [online]. Heslington, GB: Digital Preservation Coalition, 2009 [cit. 2012-01-01]. Dostupné z: http://www.dpconline.org/advice/preservationhandbook/introduction/definitions-and-concepts

DIGITAL PRESERVATION EUROPE. 2006. About DPE: mission statement. In: *DigitalPreservationEurope* [online]. 28. April 2006 [cit 2010-05-23]. Dostupné z: http://www.digitalpreservationeurope.eu/about/

DIGITAL PRESERVATION EUROPE. 2007. *DPE Research Roadmap, DPE-D7.2* [online]. Glasgow: DigitalPreservationEurope, 2007 [cit 2010-07-08]. 71 s. ISBN 978-1-906242-09-1. Dostupné z: http://www.digitalpreservationeurope.eu/publications/dpe_research_roadmap_D72.pdf

DORR, Marianne a Hartmut WEBER. 1997. *Digitization as a Means for Preservation?* [online]. Amsterdam: European Commission on Preservation and Access, 1997 [cit 2010-05-23]. Dostupné z: http://www.clir.org/pubs/reports/digpres/digpres.html

DRISCOLL, M. J. 2006. P5-MS: A general purpose tagset for manuscript description. *Digital Medievalist* [online]. 1 (2006) [cit. 2011-04-17]. ISSN 1715-0736. Dostupné z: http://www.digitalmedievalist.org/journal/2.1/driscoll/#mjd-b003

EUROPEAN COMMISSION. 2005a. *COM (2005) 465 final – Official Journal C 49 of 28.2.2008* [online]. [Brussels: European Commission], 2005 [cit. 2011-05-11]. Dostupné z: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:52005DC0465:EN:NOT

EUROPEAN COMMISSION. 2005b. Commission unveil plans for European digital libraries (IP/05/1202). In: *Europa Press Releases RAPID* [online]. Brussels: [European Commission], 30 September 2005 [cit. 2011-05-01]. Dostupné z: http://europa.eu/rapid/pressReleasesAction.do?reference=IP/05/1202&format=HTML&aged=0&language=en&guiLanguage=en

The Firenze Agenda (17 October 2003). *DigiCULT.info: a Newsletter on Digital Culture* [online]. December 2003, issue 6, s. 28-31 [cit. 2011-06-15]. ISSN 1609-3941. Dostupné z: http://www.digicult.info/downloads/dc_info_issue6_december_20031.pdf

FOJTŮ, Andrea, Jan HUTAŘ a Marek MELICHAR. 2011. Dlouhodobá ochrana digitálních dokumentů a projekt NDK. In: *Knihovny současnosti 2011: Sborník z 19. konference, konané ve dnech 13.-15. září 2011 v Českých Budějovicích.* Ostrava: Sdružení knihoven ČR, 2011, s. 73-79. ISBN 978-80-86249-62-9. Dostupné také z: http://www.svkos.cz/data/xinha/sdruk/ks2011/sbornik_2011.pdf

GANTZ, John a David REINSEL. 2011. *The 2011 Digital Universe Study: extracting Value from Chaos* [online]. IDC, 2011 [cit. 2011-12-04]. 12 s. Dostupné z: http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf

GARRETT, John a Donald WATERS. 1996. *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information* [online]. The Commission on Preservation and Access and The Research Libraries Group, 1996 [cit. 2011-11-17]. 64 s. Dostupné z: http://www.clir.org/pubs/reports/pub63watersgarrett.pdf

GAVRILIS, Dimitris, Christos PAPATHEODOROU, Panos CONSTANTOPOULOS a Stavros ANGELIS. 2010. Mopseus - A Digital Library Management System Focused on Preservation. In: LALMAS, Mounia, et al. (Eds.). *Research and Advanced Technology for Digital Libraries. 14ht European Conference, ECDL 2010 Glasgow, UK, September 6-10, 2010.* Berlin: Springer, 2010, s. 445-448. Lecture Notes in Computer Science, Vol. 6273. ISSN 0302-9743. ISBN 3-642-15463-8.

GLADNEY, Henry M. 2007. *Preserving Digital Information.* Berlin: Springer, 2007. xxxiii, 314 s. ISBN 978-3-540-37886-0.

GRADMANN, Stefan. 1998. Cataloguing vs. Metadata: old wine in new bottles? In: *64th IFLA General Conference, Amsterdam, Netherlands, August 16 - August 21, 1998* [online]. IFLA, 1998 [cit. 2011-09-16]. Dostupné z: http://archive.ifla.org/IV/ifla64/007-126e.htm

GRANGER, Stewart. 2000. Emulation as a Digital Preservation Strategy. *D-Lib Magazine* [online]. October 2000, vol. 6, nr. 10 [cit. 2011-08-07]. ISSN 1082-9873. Dostupné z: http://www.dlib.org/dlib/october00/granger/10granger.html

GREENBERG, Jane. 2005. Understanding Metadata and Metadata Schemes. In: *Metadata: a cataloger's Primer.* Binghamton (NZ): Haworth Information Press, 2005, s. 17-36. ISBN 978-0-7890-2801-3.

HANOUSEK, Tomáš. 2010. Projekt záchranné digitalizace Ad fontes v Archivu hlavního města Prahy. In: *Digitalizace and konec oslích uší: sborník konference konané v Městské knihovně v Praze 14.-16.6.2010.* Praha: Městská knihovna v Praze, 2010, s. 32-35. ISBN 978-80-85041-05-7. Dostupné také z: http://osliusi.mlp.cz/lib/exe/fetch.php/gallery/ppt_hanousek.pdf

HARGREAVES, John. [2010]. *Metadata* [online prezentace]. JISC Digital Media, [2010] [cit. 2011-10-08]. 26 snímků (ve formátu PPT). Dostupné z: http://www.slideshare.net/JISCDigi/metadata-2198853

HARVEY, Ross. 2010. *Digital curation: a how-to-do-it manual.* New York: Neal-Schuman, 2010. xxii, 225 s. How-to-do-it manuals for libraries, no. 170. ISBN 978-1-55570-694-4.

HATHI TRUST. 2011. HathiTrust Certified as Trustworthy Repository. In: *HathiTrust* [online]. 30. March 2011 [cit. 2011-04-16]. Dostupné z: http://www.hathitrust.org/hathitrust-certified-as-trustworthy-repository

HM TREASURY. 2004. *The Orange Book: Management of Risk - Principles and Concepts.* London: HM Treasury, 2004. 50 s. ISBN 1-84532-044-1. Dostupné také z: http://www.hm-treasury.gov.uk/d/orange_book.pdf

HOEVEN, Jeffrey van der a Hilde van WIJNGAARDEN. 2005. Modular emulation as a long-term preservation strategy for digital objects. In: *5th International Web Archiving Workshop (IWAW05), September 22 and 23 2005, Vienna, Austria* [online]. Alicante: European Archive, 2005 [cit. 2012-01-28]. 16 s. Dostupné z: http://iwaw.europarchive.org/05/papers/iwaw05-hoeven.pdf

HOEVEN, Jeffrey van der, Sophie SEPETJAN a Marcus DINDORF. 2010. Legal aspects of emulation. In: RAUBER, Andreas, et al. (eds.). *IPRES 2010: Proceedings of the 7th International Conference on Preservation of Digital Objects, Vienna, Austria, September 19-24, 2010.* Vienna: Oesterreichische Computer Gesellschaft, 2010, s. 113-120. ISBN 978-3-85403-262-5.

HOFMAN, Hans a Maurizio LUNGHI. 2004. Enabling Persistent And Sustainable Digital Cultural Heritage in Europe. In: *Coordinating digitisation in Europe: Progress report of the National Representatives Group: coordination mechanisms for digitisation.* Roma: MINERVA Project, 2004, s. xxx-xlv. Dostupné také z: http://www.minervaeurope.org/publications/globalreport/globalrepdf04/enabling.pdf

HOWARTH, Lynne C. 2005. Metadata and Bibliographic Control: soul-Mates or Two Solitudes? In: *Metadata: a Cataloger's Primer.* Haworth Information Press, 2005, s. 37-56. ISBN 978-0-7890-2801-3.

HUTAŘ, Jan. 2005. *METS - Metadata Encoding and Transmission Standard* [online]. Praha: [Národní knihovna ČR], 2005 [cit. 2011-09-13]. 18 s. Dostupné z: http://digit.nkp.cz/projekty/VZ-2004_2010/2005/METS_podrobne.pdf

HUTAŘ, Jan. 2008a. *Plnění administrativních metadat* [online]. Verze 1.0. Praha: Národní knihovna ČR, 2008 [cit. 2011-08-11]. 21 s. Dostupné z: http://digit.nkp.cz/Kramerius/AdminMetaData/ADMnarodniStandardVerze1_Zapis.pdf

HUTAŘ, Jan. 2008b. Proč jsou české digitální repozitáře "nespolehlivé"? *Knihovna* [online]. 2008, roč. 19, č. 2, s. 39-53 [cit. 2011-10-04]. ISSN 1801-3252. Dostupné z: http://knihovna.nkp.cz/knihovna82/82039.htm

HUTAŘ, Jan. 2008c. *Úvod do ochrany digitálních dat* [elektronická publikace]. Praha: Univerzita Karlova, Filosofická fakulta, ÚISK, 2008. 17 s.

HUTAŘ, Jan. 2011a. *Definice metadatových formátů pro digitalizaci monografií* [online]. Verze 3.0. Praha: Národní knihovna, 24. 11. 2011 [cit. 2012-02-02]. 81 s. Dostupné z: http://ndk.cz/digitalizace/nove-standardy-digitalizace-od-roku-2011/metadata_monografie_0.3

HUTAŘ, Jan. 2011b. *Definice metadatových formátů pro digitalizaci periodik* [online]. Verze 1.2. Praha: Národní knihovna, 24. 11. 2011 [cit. 2012-02-02]. 87 s. Dostupné z: http://ndk.cz/digitalizace/nove-standardy-digitalizace-od-roku-2011/metadata_periodika_v1-2_25.11

HUTAŘ, Jan a Anna NERGLOVÁ. 2007. Projekt DigitalPreservationEurope. *Čtenář.* 2007, roč. 59, č. 7-8, s. 203-208. Dostupné také z: http://ctenar.svkkl.cz/files/pdf_2007/c0707.pdf

HUTAŘ, Jan, Andrea FOJTŮ a Eliška PAVLÁSKOVÁ. 2008. DRAMBORA - nástroj na interní audit digitálních úložišť v nové online verzi a postřehy z provedených auditů. In: *Inforum 2008, Praha 28.-30.5.2008* [online]. Praha: Albertina Icome, 2008 [cit. 2011-11-14]. 9 s. ISSN 1801-2213. Dostupné z: http://www.inforum.cz/pdf/2008/hutar-jan-cze.pdf

HUTAŘ, Jan a Colin ROSENTHAL. 2008. Practical ways to tackle digital preservation using DPE tools and services. In: *WePreserve Forum, Friday 17.10 Prague* [online]. 2008 [cit. 2011-11-03]. Dostupné z: http://www.digitalpreservationeurope.eu/platter/platter_presentation_prague.pdf

HUTAŘ, Jan a Marek MELICHAR. 2010. *Návrh opatření pro prodloužení životnosti obsahu CD a DVD disků ve sbírkách NK* [online]. Praha: Národní knihovna ČR 2010 [cit. 2011-05-06]. 10 s. Dostupné z: http://digit.nkp.cz/projekty/VZ-2004_2010/2010/PDF/10_cd-archive_END.pdf

CHAPMAN, Ann, Michael DAY a Debra HIOM. 1998. Metadata: cataloguing practice and Internet subject-based information gateways. *Ariadne* [online]. December 1998, no. 18 [cit. 2011-09-16]. ISSN 1361-3200. Dostupné z: http://www.ariadne.ac.uk/issue18/metadata/

IKAROS, REDAKCE. 2005. Manuscriptorium: příprava dat a využívání informace. *Ikaros* [online]. 2005, roč. 9, č. 5/2 [cit. 2011-11-10]. URN-NBN:cz-ik1904. ISSN 1212-5075. Dostupné z: http://www.ikaros.cz/node/1904

Implementace formátu METS v Systému Kramerius. In: *Informační systém výzkumu, experimentálního vývoje a inovací* [online]. 2007 [cit. 2011-04-25]. Dostupné z: http://www.isvav.cz/resultDetail.do?rowId=RIV/00023221:_____/07:%230000025!RIV07-MK0-00023221

ISO/IEC 21000-2:2003 *- Information technology - Multimedia framework (MPEG-21) - Part 2: Digital Item Declaration*. 1st edition. Geneve: ISO, 2005. 88 s.

JONÁK, Zdeněk. 2003. Data. In: *KTD: Česká terminologická databáze knihovnictví a informační vědy (TDKIV)* [online]. Praha: Národní knihovna ČR, 2003 [cit. 2011-09-16]. Sysno 000000442. Dostupné z: http://aleph.nkp.cz/F/?func=direct&doc_number=000000442&local_base=KTD

KAHLE, Brewster. 2004. Towards universal access to all knowledge. In: *Strategies for a european area of digital cultural resources: towards a continuum of digital heritage. European conference, The Hague, The Netherlands 15-16 September 2004.* Haag, 2004, s. 27-34.

KAHN, Robert a Robert WILENSKI. 1995. *A Framework for Distributed Digital Object Services* [online]. [Reston: Corporation for National Research Initiatives], 13. květen 1995 [cit. 2011-03-13]. Dostupné z: http://www.cnri.reston.va.us/home/cstr/arch/k-w.html

KAŠPAROVÁ, Jaroslava a Tomáš PSOHLAVEC. 2008. *Interoperabitita TEI P5/MARC 21: katalogizace historických dokumentů* [online]. Praha: Národní knihovna ČR, 2008 [cit. 2011-06-17]. 190 s. Dostupné z: http://digit.nkp.cz/projekty/VZ-2004_2010/2008/Prilohy/Interoperabilita_VaV_v011.pdf

KIM, Yunhyong a Seamus ROSS. 2006. Genre Classification in Automated Ingest and Appraisal Metadata. In: GONZALO, Julio, et al. *(Eds.). Research and Advanced Technology for Digital Libraries: 10th European Conference, ECDL 2006 Alicante, Spain, September 17-22, 2006.* Berlin: Springer, 2006, s. 63-74. Lecture Notes in Computer Science, Vol. 4172. ISSN 0302-9743. ISBN 3-540-44636-2.

KIRCHHOFF, Amy, et al. 2010. Becoming a certified trustworthy digital repository: the Portico experience. In: RAUBER, Andreas, et al. (eds.). *IPRES 2010: Proceedings of the 7th International Conference on Preservation of Digital Objects, Vienna, Austria, September 19-24, 2010.* Vienna: Oesterreichische Computer Gesellschaft, 2010, s. 87-93. ISBN 978-3-85403-262-5.

KNIGHT, Steve. 2011. *Osobní rozhovor na konferenci iPRES2011 v Singapuru*. Singapore, 2011.

KNOLL, Adolf. 1997. Elektronické publikace v Národní knihovně České republiky. In: *Automatizace knihovnických procesů 1997*, s. 11-21. Dostupné také z: http://digit.nkp.cz/CzechArticles/Elpubl_katedra.html

KNOLL, Adolf. 1998. Memoriae Mundi Series Bohemica: Program digitálního zpřístupnění vzácných fondů. *Ikaros* [online]. 1998, roč. 2, č. 7 [cit. 2011-12-02]. URN-NBN:cz-ik1175. ISSN 1212-5075. Dostupné z: http://www.ikaros.cz/node/1175

KNOLL, Adolf. 1999. Problematika elektronických publikací. *Národní knihovna: knihovnická revue.* 1999, roč. 10, č. 4, s. 173-177. ISSN 1214-0678. Dostupné také z: http://knihovna.nkp.cz/Nkkr9904/9904173.html

KNOLL, Adolf. 2000. *Technical Standards Used by the National Digitization Programmes Initiated by the National Library of the Czech Republic* [online]. Praha: Národní knihovna ČR, 2000 [cit. 2011-07-08]. 8 s. Dostupné z: http://digit.nkp.cz/EnglishArticles/Technical_Standards.rtf

KNOLL, Adolf. 2002. Technical Standards Used by the National Digitization Programmes Initiated by the National Library of the Czech Republic in Comparison with DIEPER. In: *CASLIN 2002: Ochrana a sprístupňovanie dokumentov: nové trendy: Zborník z konferencie, konanej 23.- 27. júna 2000 v Pribyline.* Martin: Slovenská národná knižnica, 2002, s. 37-45.

KNOLL, Adolf. 2003. Digital Library for Access to Rare Materials: from pilot projects to national digitisation programmes. *LIBER Quarterly.* 2003, Vol. 13, No. 3/4, s. 232-240. Dostupné také z: http://liber.library.uu.nl/publish/articles/000030/article.pdf

KNOLL, Adolf. 2004. *Digitální knihovna - produkce, ochrana a zpřístupnění digitálních dokumentů: závěrečná zpráva o řešení výzkumného záměru* [online]. Praha: Národní knihovna ČR, 2004 [cit. 2011-09-04]. 15 s. Dostupné z: http://digit.nkp.cz/PROJECTS_WEB/CEZFinal1999-2003.pdf

KNOLL, Adolf. 2009. Projekt ENRICH: budujeme Evropskou digitální knihovnu rukopisů. In: *Mezinárodní muzeologická konference Muzeum a změna III* [online]. 19. únor 2009 [cit. 2011-11-12]. Dostupné z: http://www.cz-museums.cz/amg/UserFiles/File/muchang%20III/knoll_muchang.doc

KNOLL, Adolf. 2010a. Ohlédnutí za digitalizací v českých knihovnách: jak to bylo. In: *Digitalizace and konec oslích uší. Sborník konference konané v Městské knihovně v Praze 14.-16.6.2010.* Praha: Městská knihovna v Praze, 2010, s. 20-30. ISBN 978-80-85041-05-7.

KNOLL, Adolf. 2010b. *Vytvoření virtuálního badatelského prostředí pro zpřístupnění a ochranu digitálních dokumentů: dílčí zpráva za r. 2010.* Praha: Národní knihovna ČR, 2010. 38 s. Dostupné také z: http://digit.nkp.cz/projekty/VZ-2004_2010/2010/ZpravaInet.pdf

KNOLL, Adolf. 2011a. *Emailová korespondence ze dne 20. 4. 2011* [online]. [cit. 2011-05-05].

KNOLL, Adolf. 2011b. *VISK 6 - Národní program digitálního zpřístupnění vzácných dokumentů Memoriae Mundi Series Bohemica* [online]. 14. červen 2011 [cit. 2011-05-23]. Dostupné z: http://visk.nkp.cz/VISK6.htm

KNOLL, Adolf a Tomáš MAYER. 1999. The Structure of Digital Copies of Old Books and Manuscripts II. In: *Digitization of Rare Library Materials: storage and access to data* [CD-ROM]. Version 2.1. Praha: Národní knihovna ČR, 1999. Dostupné také z: http://www.unesco.org/webworld/mdm/czech_digitization/doc/digit_2.htm

KNOLL, Adolf a Stanislav PSOHLAVEC. 1999. *Digitization of Rare Library Materials: storage and access to data* [CD-ROM]. Version 2.1. Praha: Národní knihovna ČR, 1999. Dostupné také z: http://www.unesco.org/webworld/mdm/czech_digitization/doc/lev2.htm

KNOLL, Adolf a Jan VOMLEL. 1999a. Digitization of Old Books, Manuscripts and other Documents: the Structure for Metadata for Digital Copies of Books. In: *Digitization of Rare Library Materials: storage and access to data* [CD-ROM]. Version 2.1. Praha: Národní knihovna ČR, 1999. Dostupné také z: http://www.unesco.org/webworld/mdm/czech_digitization/doc/mod_book.htm

KNOLL, Adolf a Jan VOMLEL. 1999b. The Format for Storage of Metadata. In: *Digitization of Rare Library Materials: storage and access to data* [CD-ROM]. Version 2.1. Praha: Národní knihovna ČR, 1999. Dostupné také z: http://www.unesco.org/webworld/mdm/czech_digitization/doc/digitiz.htm

KNOLL, Adolf a Jan VOMLEL. 1999c. Metadata Structures for Digital Copies of Periodicals. In: *Digitization of Rare Library Materials: storage and access to data* [CD-ROM]. Version 2.1. Praha:

Národní knihovna ČR, 1999. Dostupné také z: http://www.unesco.org/webworld/mdm/czech_digitization/doc/periodic.htm

KNOLL, Adolf a Stanislav PSOHLAVEC. 2002. *Zpráva o řešení projektu výzkumu a vývoje Optimalizace archivace a zpřístupnění digitálních dat: závěrečná zpráva za léta 2001-2002.* Praha: Národní knihovna ČR, 2002. 20 s. Dostupné z: http://digit.nkp.cz/knihcin/digit/vav23/zpravaweb.doc. V názvu chybně uvedeny roky trvání projektu, má být 2000-2001.

KNOLL, Adolf, Jiří POLIŠENSKÝ a Zdeněk UHLÍŘ. 2004. *Vytvoření virtuálního badatelského prostředí pro zpřístupnění a ochranu digitálních dokumentů: Zpráva o řešení výzkumného záměru za rok 2004.* Praha: Národní knihovna ČR, 2004. 41 s. Dostupné také z: http://digit.nkp.cz/projekty/VZ-2004_2010/2004/Zprava2004.pdf

KNOLL, Adolf, Jiří POLIŠENSKÝ a Zdeněk UHLÍŘ. 2005. *Vytvoření virtuálního badatelského prostředí pro zpřístupnění a ochranu digitálních dokumentů: průběžná zpráva o řešení za rok 2005.* Praha: Národní knihovna ČR, 2005. 17 s. Dostupné také z: http://digit.nkp.cz/projekty/VZ-2004_2010/2005/Zprava2005.pdf

KNOLL, Adolf, Jiří POLIŠENSKÝ a Zdeněk UHLÍŘ. 2006. *Vytvoření virtuálního badatelského prostředí pro zpřístupnění a ochranu digitálních dokumentů: dílčí zpráva o řešení za rok 2006.* Praha: Národní knihovna ČR, 2006. 21 s. Dostupné také z: http://digit.nkp.cz/projekty/VZ-2004_2010/2006/ZPRAVA2006.pdf

KNOLL, Adolf, Jiří POLIŠENSKÝ a Zdeněk UHLÍŘ. 2007. *Vytvoření virtuálního badatelského prostředí pro zpřístupnění a ochranu digitálních dokumentů: dílčí zpráva o řešení za rok 2007.* Praha: Národní knihovna ČR, 2007. 21 s. Dostupné také z: http://digit.nkp.cz/projekty/VZ-2004_2010/2007/ZpravaInternet.pdf

KNOLL, Adolf, Jiří POLIŠENSKÝ a Zdeněk UHLÍŘ. 2008. *Vytvoření virtuálního badatelského prostředí pro zpřístupnění a ochranu digitálních dokumentů: dílčí zpráva o řešení za rok 2008.* Praha: Národní knihovna ČR, 2008. 21 s. Dostupné také z: http://digit.nkp.cz/projekty/VZ-2004_2010/2008/Zprava08_Inet.pdf

KNOLL, Adolf, Jiří POLIŠENSKÝ a Zdeněk UHLÍŘ. 2009. *Vytvoření virtuálního badatelského prostředí pro zpřístupnění a ochranu digitálních dokumentů: dílčí zpráva o řešení za r. 2009.* Praha: Národní knihovna ČR, 2009. 25 s. Dostupné také z: http://digit.nkp.cz/projekty/VZ-2004_2010/2009/HlavniZpravaWeb.pdf

Koncepce trvalého uchování knihovních sbírek tradičních a elektronických dokumentů v knihovnách ČR do roku 2010. *Knihovna - knihovnická revue.* 2005, roč. 16, č. 2, s. 9-27. Dostupné také z: http://knihovna.nkp.cz/pdf/0502/050209.pdf

KONSTANTELOS, Leo. 2012. *Osobní rozhovor v Národní knihovně Nového Zélandu 2.3.2012*. Wellington, 2012.

KORENKOVA, Margarita a Ann HÄGERFORS. 2011. Quality Criteria for Digital Information in Long-term Digital Preservation. In: *Archiving 2011, May 16-19, 2011, Salt Lake City, Utah. Final Program and Proceedings.* Springfield (VA): Society for Imaging Science and Technology, 2011, s. 34-39. ISBN 978-0-89208-294-0.

KRIMBACHER, Monika, Michael NEUHAUSER a Martina VOGL. 2005. *reUSE: Survey on the Long-Term Preservation of Digital Documents in European Libraries 2005.* Innsbruck: University Innsbruck Library, 2005. 103 s.

KUČERA, Karel a Stanislav PSOHLAVEC. 2001. *DTD pro projekt Memoriae mundi series Bohemica: s návazností na projekt MASTER TEI a s implementací standardizovaných technických informací o obrazech.* Beroun: AiP, 2001. 75 s. Dostupné také z: http://digit.nkp.cz/knihcin/digit/vav23/MSSDTD.pdf

KUNT, Miroslav. 2006. Dlouhodobé ukládání elektronických dokumentů v novém archivním zákonu. In: *Archivy, knihovny a muzea v digitálním světě 2005.* Praha: Národní technické muzeum, 2006, s. 2-35. ISBN 80-7037-149-8.

KUNY, Terry. 1998. The Digital Dark Ages? Challenges In The Preservation Of Electronic Information. *International Preservation News* [online]. May 1998, no. 17 [cit. 2011-09-11]. ISSN 0890-4960. Dostupné z: http://archive.ifla.org/VI/4/news/17-98.htm

LAGOZE, Carl. 1996. The Warwick Framework: a Container Architecture for Diverse Sets of Metadata. *D-Lib Magazine* [online]. July/August 1996, Vol. 2, Issue 7/8 [cit. 2011-12-14]. ISSN 1082-9873. Dostupné z: http://www.dlib.org/dlib/july96/lagoze/07lagoze.html

LAVOIE, Brian. 2004. *The Open Archival Information System Reference Model: Introductory Guide: Technology Watch Report* [online]. Dublin (OH): OCLC & DPC, 2004 [cit. 2011-11-08]. 20 s. DPC Technology Watch Series Report 04-01. Dostupné z: http://www.dpconline.org/docs/lavoie_OAIS.pdf

LAVOIE, Brian a Lorcan DEMPSEY. 2004. Thirteen Ways of Looking at... Digital Preservation. *D-Lib Magazine* [online]. July/August 2004, Volume 10, Number 7/8 [cit. 2011-12-29]. ISSN 1082-9873. DOI 10.1045/july2004-lavoie. Dostupné z: http://www.dlib.org/dlib/july04/lavoie/07lavoie.html

LAVOIE, Brian a Richard GARTNER. 2005. *Preservation metadata: Technology Watch Report* [online]. [Dublin (OH)]: OCLC, DPC, Oxford University Library Services, 2005 [cit. 2011-08-14]. 21 s. DPC Technology Watch Series Report 05-01. Dostupné z: http://www.dpconline.org/docs/reports/dpctw05-01.pdf

LAWRENCE, Gregory W., et al. 2000. *Risk Management of Digital Information: a File Format Investigation.* Washington (DC): Council on Library and Information Resources, 2000. 75 s. ISBN 1-887334-78-5. Dostupné také z: http://www.clir.org/pubs/reports/pub93/pub93.pdf

LESK, Michael. 1992. Preservation of New Technology: a report of the Technology Assessment Advisory Committee to the Commission on Preservation and Access. In: *Council on Library and Information resources* [online]. 1992 [cit. 2011-08-21]. Dostupné z: http://www.clir.org/pubs/reports/lesk/lesk2.html

LHOTÁK, Martin. 2007. Open source pro digitální knihovnu. In: *Automatizace knihovnických procesů – 11: sborník z 11. ročníku semináře pořádaného ve dnech 16.–17. května 2007 v Liberci.* Praha: ČVUT, 2007, s. 65-72. ISBN 978-80-01-03691-4. Dostupné také z: http://www.akvs.cz/akp-2007/09-lhotak.pdf

LJUBKA, Ivan. 2008. Kramerius - vývoj aplikace pro zpřístupnění. In: *Sborník z 16. konference, konané ve dnech 16.-18. září 2008 v Seči u Chrudimi.* Brno: Sdružení knihoven ČR, 2008, s. 91-94. ISBN 978-80-86249-49-0. Dostupné také z: http://www.svkos.cz/data/xinha/sdruk/2008-1-091.pdf

LOCKSS ALLIANCE. [2008]. About Us – LOCKSS. *Home - LOCKSS* [online]. Palo Alto (CA): LOCKSS, [2008] [cit. 2011-12-03]. Dostupné z: http://www.lockss.org/lockss/About_Us

LUNGHI, Maurizio. 2004. European Actions for Sustainability and Preservation. In: *Strategies for a european area of digital cultural resources: towards a continuum of digital heritage: European conference, The Hague, The Netherlands 15-16 September 2004.* Haag: [s.n.], 2004, s. 79-86.

LUPOVICI, Catherine a Julien MASANÈS. 2000. *Metadata for long term-preservation* [online]. Paris: NEDLIB Consortium, 2000 [cit. 2011-10-04]. 25 s. Dostupné z: http://www.kb.nl/hrd/dd/dd_links_en_publicaties/nedlib/preservationmetadata.pdf

McDONOUGH, Jerome. 2003. METS: a Status Report. In: *Project Briefing: Spring 2003 Task Force Meeting 28-29 April 2003* [online]. Coalition for Networked Information, 2003 [cit. 2011-12-18]. 31 snímků (ve formátu PPT). Dostupné z: http://old.cni.org/tfms/2003a.spring/powerpoints/PPT-METS-McDonough.ppt

MELICHAR, Marek a Jan HUTAŘ. 2011. *Specifikace požadavků na dodávku technologií a služeb*. Praha: Národní knihovna ČR, 2011. 106 s. Interní podklad k tendru NDK projektu ze 4.2.2011.

MILSTEAD, Jessica a Susan FELDMAN. 1999. Metadata: cataloging by Any Other Name. *Online* [online]. January/February 1999, Vol. 23, Nr. 1, s. 24-26, 28-31 [cit. 2011-12-15]. ISSN-0146-5422. Dostupné z: http://www.highbeam.com/doc/1G1-53457500.html

NÁRODNÍ ARCHIV. 2009. *Závěrečná zpráva projektu výzkumu a vývoje* [online]. Praha: Národní archiv ČR, 2009 [cit. 2012-01-29]. 14 s. Dostupné z: http://www.nacr.cz/Z-files/moznosti_01.pdf

NÁRODNÍ KNIHOVNA ČR. 2007. *Pravidla popisu pro tvorbu metadat periodik a monografií.* Praha: Národní knihovna ČR, 2007. 25 s. Dostupné také z: http://digit.nkp.cz/projekty/VZ-2004_2010/2007/Prilohy/3.pdf

NÁRODNÍ KNIHOVNA ČR. 2008. *Pravidla popisu periodik verze 6.4.* Praha: Národní knihovna ČR, 2008. 19 s. Dostupné také z: http://kramerius-info.nkp.cz/visk/zadavaci-dokumentace-visk-7-pro-rok-2010/pravidla-popisu-periodik

NÁRODNÍ KNIHOVNA ČR. 2009. *Pravidla popisu monografií verze 1.5.* Praha: Národní knihovna ČR, 2009. 10 s. Dostupné také z: http://kramerius-info.nkp.cz/visk/zadavaci-dokumentace-visk-7-pro-rok-2010/pravidla-popisu-monografii

NÁRODNÍ KNIHOVNA ČR. 2011. Dotační projekt vytvoření Národní digitální knihovny: výběrové řízení na systémového integrátora: zadávací dokumentace. In: *Národní digitální knihovna* [online]. Praha: Národní knihovna, 12. 12. 2011(poslední změna) [cit. 2012-01-14]. Dostupné z: http://ndk.cz/narodni-dk/dotacni-projekt-vytvoreni-narodni-digitalni-knihovny/vyberove-rizeni-na-systemoveho-integratora/zadavaci-dokumentace/jednotlive-dokumenty

NÁRODNÍ TECHNICKÁ KNIHOVNA. 2011. *Audit důvěryhodnosti Digitálního repozitáře NUŠL* [online]. 8. únor 2011 [cit. 2011-05-01]. Dostupné z: http://nusl.techlib.cz/index.php/Audit

NATIONAL LIBRARY OF AUSTRALIA. 2003. *Guidelines for the Preservation of Digital Heritage* [online]. Paris: UNESCO, Information Society Division, 2003 [cit. 2011-09-23]. 170 s. Dostupné z: http://unesdoc.unesco.org/images/0013/001300/130071e.pdf

NATIONAL LIBRARY OF NEW ZEALAND. 2003. *Metadata Standards Framework: Preservation Metadata (Revised)* [online]. Wellington: National Library of New Zealand, 2003 [cit. 2011-09-23]. 51 s. Dostupné z: http://www.natlib.govt.nz/downloads/metaschema-revised.pdf

NESTOR. 2008. *Kriterienkatalog vertrauenswürdige digitale Langzeitarchive* [online]. Frankfurt am Main: Nestor Working Group, 2008 [cit. 2011-08-21]. 55 s. urn:nbn:de:0008-2008021802. Dostupné z: http://edoc.hu-berlin.de/series/nestor-materialien/8/PDF/8.pdf

NESTOR. 2009. *Catalogue of Criteria for Trusted Digital Repositories* [online]. Version 2. Frankfurt am Main: nestor c/o Deutsche Nationalbibliothek, 2009 [cit. 2011-02-27]. 53 s. Nestor materials, 8. urn:nbn:de:0008-2010030806. Dostupné z: http://files.d-nb.de/nestor/materialien/nestor_mat_08_eng.pdf

NISO. 2004. *Understanting Metadata* [online]. Bethesda (MD): National Information Standards Organization Press, 2004 [cit. 2011-09-18]. 16 s. ISBN 1-880124-62-9. Dostupné z: http://www.niso.org/publications/press/UnderstandingMetadata.pdf

NOVOTNÝ, Jan. 2011. Restaurátorský informační systém ResIS. In: *Výzkum a vývoj nových postupů v ochraně a konzervaci písemných památek 2005-2011.* Praha: Národní knihovna ČR, 2011, s. 93-116. ISSN 978-80-7050-603-5.

OCLC. 2002. *Digital Archive Metadata Elements: OCLC Digital Archive System Guides* [online]. Dublin (OH): OCLC, 2002 [cit. 2011-09-23]. 44 s. Dostupné z:

http://www.oclc.org/support/documentation/digitalarchive/da_metadata_elements/da_metadata_elements.pdf

OCLC/CRL. 2007. *Trustworthy Repositories Audit & Certification: Criteria and Checklist* [online]. Version 1.0. Dublin (OH): OCLC, February 2007 [cit. 2011-09-12]. 88 s. Dostupné z: http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf

OCLC/RLG PREMIS WORKING GROUP. 2004. *Implementing Preservation Repositories for Digital Materials: Current Practice And Emerging Trends In The Cultural Heritage Community: Report by the joint OCLC/RLG Working Group Preservation Metadata: Implementation Strategies (PREMIS)* [online]. Dublin (OH): OCLC, September 2004 [cit. 2011-05-13]. 66 s. Dostupné z: http://www.oclc.org/research/activities/past/orprojects/pmwg/surveyreport.pdf

OCLC/RLG PREMIS WORKING GROUP. 2005. *Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group* [online]. Dublin (OH): OCLC, May 2005 [cit. 2011-08-21]. 237 s. Dostupné z: http://www.oclc.org/research/activities/past/orprojects/pmwg/premis-final.pdf

OCLC/RLG WORKING GROUP ON PRESERVATION METADATA. 2001. *Preservation Metadata for Digital Objects: a Review of the State of the Art A White Paper by the OCLC/RLG Working Group on Preservation Metadata* [online]. Dublin (OH): OCLC, RLG, 31. January 2001 [cit. 2011-09-25]. 50 s. Dostupné z: http://www.oclc.org/research/activities/past/orprojects/pmwg/presmeta_wp.pdf

OCLC/RLG WORKING GROUP ON PRESERVATION METADATA. 2002. *Preservation Metadata and the OAIS Information Model: a Metadata Framework to Support the Preservation of Digital Objects* [online]. Dublin (OH): OCLC, June 2002 [cit. 2011-10-08]. 54 s. Dostupné z: http://www.oclc.org/research/activities/past/orprojects/pmwg/pm_framework.pdf

PALFREY, John G. a Urs GASSER. 2008. *Born digital: understanding the first generation of digital natives.* New York: Basic Books, 2008. 288 s. ISBN 9780465005154.

PAVLICOVÁ, Libuše. 2001. Systém CDS/ISIS-MAKS v Knihovně B.B. Buchlovana. *Knihovnicko-informační zpravodaj U nás* [online]. 2001, roč. 11, č. 1, s. 1-3 [cit. 2011-09-09]. Dostupné z: http://www.svkhk.cz/SVKHK/u-nas-pdf_archiv/419.pdf

POLIŠENSKÝ, Jiří. 2000. Hybridní technologie reformátování knihovních fondů. In: *Inforum 2000. 23-25.5.2000, Praha* [online]. Praha: AiP Beroun, 2000 [cit. 2011-12-02]. Dostupné z: http://www.inforum.cz/archiv/inforum2000/prednasky/hybridnitechno.htm

POLIŠENSKÝ, Jiří. 2002. Konsorcium pro zpřístupňování a archivaci dokumentů pro výzkum a vývoj v České republice. In: *CASLIN 2002. Ochrana a sprístupňovanie dokumentov: nové trendy: Zborník z konferencie, konanej 23.- 27. júna 2000 v Pribyline.* Martin: Slovenská národná knižnica, 2002, s. 57-63.

POLIŠENSKÝ, Jiří. 2004. Role Systému Kramerius v oblasti tvorby a zpřístupňování digitálních dokumentů. *Knihovnický zpravodaj Vysočina* [online]. 2004, roč. 4, č. 1 [cit. 2011-09-10]. ISSN 1213-8231. Dostupné z: http://kzv.kkvysociny.cz/archiv.aspx?id=391&idr=2&idci=6

POLIŠENSKÝ, Jiří. 2007. Implementace formátu METS v Systému Kramerius. In: *Automatizace knihovnických procesů – 11: sborník z 11. ročníku semináře pořádaného ve dnech 16.–17. května 2007 v Liberci* [online]. Praha: ČVUT, 2007 [cit. 2011-08-08]. ISBN 978-80-01-03691-4. Dostupné z: http://www.akvs.cz/akp-2007/13-polisensky.pdf

POLIŠENSKÝ, Jiří. 2008. Současné trendy v digitalizaci novodobých dokumentů. In: *Sborník z 16. konference, konané ve dnech 16.-18. září 2008 v Seči u Chrudimi* [online]. Brno: Sdružení knihoven ČR, 2008 [cit. 2011-09-10]. s. 51-60. ISBN 978-80-86249-49-0. Dostupné z: http://www.svkos.cz/data/xinha/sdruk/2008-1-051.pdf

PREMIS EDITORIAL COMMITTEE. 2011. *PREMIS Data Dictionary for Preservation Metadata* [online]. Version 2.1. Washington: PREMIS Editorial Committee, January 2011 [cit. 2011-10-14]. 226 s. Dostupné z: http://www.loc.gov/standards/premis/v2/premis-2-1.pdf

PRENSKY, Marc. 2001. Digital Natives, Digital Immigrants. *On the Horizon* [online]. 2001, Vol. 9, No. 5, 15 s. [cit. 2011-09-17]. Dostupné z: http://www.albertomattiacci.it/docs/did/Digital_Natives_Digital_Immigrants.pdf

PSOHLAVEC, Stanislav. 2004. Projekt MEMORIA, rukopisy a staré tisky na internetu. *Národní knihovna: knihovnická revue* [online]. 2004, roč. 15, č. 1, s. 40-43 [cit. 2011-10-23]. ISSN 1214-0678. Dostupné z: http://knihovna.nkp.cz/NKKR0401/0401040.html

PSOHLAVEC, Stanislav. 2006. Možnosti intenzivní digitalizace časopisů a knih. In: *Archivy, knihovny a muzea v digitálním světě 2005.* Praha: Národní technické muzeum v Praze, 2006, s. 39-42. ISBN 80-7037-149-8.

PSOHLAVEC, Tomáš. 2009. *Pokračování vývoje M-Toolu* [online]. Verze 1.0. Beroun: AiP Beroun, 2009 [cit. 2011-03-20]. 78 s. Dostupné z: http://digit.nkp.cz/projekty/VZ-2004_2010/2009/Prilohy/2_MNS_M-TOOL_Cont.pdf

QBIZM TECHNOLOGIES. 2007. *Systém pro zpřístupnění digitálních dokumentů Kramerius*. Verze: 8 final – 2.5.2007. Brno, 2007. 88 s. Interní dokument NK ČR ve formátu MS Word.

RÁKOCY, Veronika. 2008. *Zpřístupňování písemného kulturního dědictví: projekty Manuscriptorium a Manuscriptorium pro školy.* Brno, 2008. 110 s., 2 s. příl. Diplomová práce. Masarykova univerzita, Filozofická fakulta, Kabinet informačních studií a knihovnictví. Vedoucí diplomové práce Zdeněk Uhlíř. Dostupné také z: http://is.muni.cz/th/64460/

RAS, Marcel. 2009. *Osobní rozhovor při návštěvě KB NL*. Haag, 9. červen 2009.

REESE, Terry. 2008. *Building digital libraries: a how-to-do-it manual.* New York (NY): Neal-Schuman, 2008. xv, 277 s. How-to-do-it manuals for libraries, no. 153. ISBN 978-1-55570-617-3.

RESEARCH LIBRARIES GROUP. 1998. RLG Working Group on Preservation Issues of Metadata: final Report. In: *Research Libraries Group* [online]. Research Libraries Group, 21. May 1998 [cit. 2011-09-23]. Dostupné z: http://web.archive.org/web/19981205114312/http://www.rlg.org/preserv/presmeta.html. Zdroj je dostupný pouze v Internet Archive.

RESEARCH LIBRARIES GROUP. 2002. *Trusted Digital Repositories: Attributes and Responsibilities: An RLG-OCLC Report* [online]. Mountain View (CA): RLG, 2002 [cit. 2011-07-25]. vi, 62 s. Dostupné z: http://www.oclc.org/research/activities/past/rlg/trustedrep/repositories.pdf

ROSENTHAL, Colin, Asger BLEKINGE-RASMUSSEN a Jan HUTAŘ. 2009. *Průvodce plánem důvěryhodného digitálního repozitáře (PLATTER).* Praha: Národní knihovna ČR, 2009. 51 s. ISBN 978-80-7050-569-4.

ROSS, Seamus. 2000. *Changing Trains at Wigan: Digital Preservation and the Future of Scholarship*. London: National Preservation Office, 2000. 44 s. Dostupné také z: http://www.bl.uk/blpac/pdf/wigan.pdf

ROSS, Seamus. 2004. Reflections on the Impact of the Lund Principles on European Approaches to Digitisation. In: *Strategies for a european area of digital cultural resources: towards a continuum of digital heritage. European conference, The Hague, The Netherlands 15-16 September 2004.* Haag, 2004, s. 87-98.

ROSS, Seamus. 2011. Digital Preservation: Why should today's society pay for the benefit of society in future. In: *iPRES 2011. 8th International Conference on Preservation of Digital Objects, 1.- 4. 11. 2011, Singapore.* Keynote přednáška. Abstrakt dostupný z: http://ipres2011.sg/pages/keynotes

ROSS, Seamus, et al. 2009. *An Introduction to the DRAMBORA toolkit and its Underlying Principles* [online prezentace]. Barcelona, 26. 3. 2009 [cit. 2011-04-17]. 116 snímků (ve formátu PPT). Dostupné z: http://www.digitalpreservationeurope.eu/preservation-training-materials/files/DRAMBORA_barcelona.ppt

ROTHENBERG, Jeff. 1999. *Avoiding technological quicksand: finding a viable technical foundation for digital preservation.* Washington (DC): Council on Library and Information Resources, 1999. 35 s. ISBN 1-887334-63-7. Dostupné také z: http://www.clir.org/pubs/reports/rothenberg/pub77.pdf

RUSSELL, Kelly. 1998. EDARS: Long-term Access and Usability of Digital Resources: the Digital Preservation Conundrum. *Ariadne* [online]. December 1998, no. 18 [cit. 2011-09-16]. ISSN 1361-3200. Dostupné z: http://www.ariadne.ac.uk/issue18/cedars/

RUSSELL, Kelly, et al. 2000. *Metadata For Digital Preservation: The Cedars Project Outline Specification* [online]. CEDARS, March 2000 [cit. 2011-09-24]. 33 s. Dostupné z: http://web.archive.org/web/20010613073650/http://www.leeds.ac.uk/cedars/MD-STR~5.pdf. Zdroj je dostupný pouze v Internet Archive.

SHENTON, Helen. [2004]. Digital versus print as a preservation format - expert views from international comparator libraries. In: *The British Library - The world's knowledge* [online]. London: British Library, [2004] [cit. 2012-01-06]. Dostupné z: http://www.bl.uk/aboutus/stratpolprog/ccare/introduction/digital/digitalvprint/index.html

SCHREIBMAN, Susan (ed.). 2007. *Best Practice Guidelines for Digital Collections* [online]. College Park (MD), May 2007 [cit. 2011-10-08]. 76 s. Dostupné z: http://www.lib.umd.edu/dcr/publications/best_practice.pdf

SCHULTZ, Matt a Emily B. GORE. 2010. The importance of trust in distributed digital preservation: a case study from the MetaArchive cooperative. In: RAUBER, Andreas, et al. (eds.). *IPRES 2010: Proceedings of the 7th International Conference on Preservation of Digital Objects, Vienna, Austria, September 19-24, 2010.* Vienna: Oesterreichische Computer Gesellschaft, 2010, s. 105-111. ISBN 978-3-85403-262-5.

SINCLAIR, Pauline a Amir BERNSTEIN. 2010. *An Emerging Market: Establishing Demand for Digital Preservation Tools and Services* [online]. PLANETS, 2010 [cit. 2011-12-09]. 10 s. Planets White Paper. Dostupné z: http://www.planets-project.eu/docs/reports/Planets-VENDOR-White-Paperv4.pdf

SKLENÁK, Vilém. 2003. SGML. In: *KTD: Česká terminologická databáze knihovnictví a informační vědy (TDKIV)* [online]. Praha: Národní knihovna ČR, 2003 [cit. 2011-09-16]. Sysno 000000658. Dostupné z: http://aleph.nkp.cz/F/?func=direct&doc_number=000000658&local_base=KTD

SMIRAGLIA, Richard P. 2005. Introducing Metadata. In: *Metadata: a cataloger's Primer.* Binghamton (NY): Haworth Information Press, 2005, s. 1-15. ISBN 978-0-7890-2801-3.

SMITH, Abby. 1999. *Why Digitize?* Washington (DC): Council on Library and Information Resources, 1999. 13 s. ISBN 1-887334-65-3. Dostupné také z: http://www.clir.org/pubs/reports/pub80-smith/pub80.pdf

SMITH, Abby. 2002. Rethinking the Library in Digital Age. In: *CASLIN 2002. Ochrana a sprístupňovanie dokumentov: nové trendy: Zborník z konferencie, konanej 23.- 27. júna 2000 v Pribyline.* Martin: Slovenská národná knižnica, 2002, s. 5-12

STEINKE, Tobias. 2005. *LMER: Long-term preservation Metadata for Electronic Resources* [online]. Frankfurt a. Main: Deutsche Bibliothek, 2005 [cit. 2011-09-20]. 19 s. urn:nbn:de:1111-2005051906. Dostupné z: http://www.d-nb.de/standards/pdf/lmer12_e.pdf

STEINKE, Tobias. 2009. DNB, LMER. In: *Deutsche Nationalbibliothek* [online]. [Leipzig]: Deutsche Nationalbibliothek, 12. März 2009 [cit. 2011-09-20]. Dostupné z: http://www.d-nb.de/eng/standards/lmer/lmer.htm

STEINKE, Tobias. 2011. *Osobní rozhovor, IIPC meeting.* Londýn, 7. 10 2011.

STOKLASOVÁ, Bohdana. 2001. Přežije formát UNIMARC rok 2003? *Ikaros* [online]. 2001, roč. 5, č. 9 [cit. 2011-08-22]. ISSN 1212-5075. Dostupné z: http://www.ikaros.cz/node/815

STOKLASOVÁ, Bohdana. 2004. Formát MARC21: čím se liší od formátu UNIMARC. In: *Národní knihovna České republiky* [online]. 9. leden 2004 [cit. 2011-08-22]. Dostupné z: http://www.nkp.cz/pages/page.php3?page=fond_Marc21.htm

STOKLASOVÁ, Bohdana. 2006. Teze Koncepce trvalého uchování knihovních sbírek tradičních a elektronických dokumentů v knihovnách ČR do roku 2010. In: *Archivy, knihovny a muzea v digitálním světě 2005.* Praha: Národní technické muzeum, 2006, s. 108-113. ISBN 80-7037-149-8.

STOKLASOVÁ, Bohdana a Martin SVOBODA. 1991. Computerization of Czech Libraries: History, present and future. In: *[International Conference on New Information Technology].* [Budapest], 1991, s. [217-223]. Dostupné také z: http://web.simmons.edu/~chen/nit/NIT'91/217-svo.htm

STOKLASOVÁ, Bohdana a Jan HUTAŘ. 2007. Nové směry v dlouhodobém uchovávání dokumentů v mezinárodním kontextu. In: *Automatizace knihovnických procesů 11. Liberec 16.-17.5.2007.* Praha: ČVUT, 2007, s. 87-96. ISBN 978-80-01-03691-4

STRODL, Stephan, et al. 2010. Automating Logical Preservation for Small Institutions with Hoppla. In: LALMAS, Mounia, et al. (Eds.). *Research and Advanced Technology for Digital Libraries. 14ht European Conference, ECDL 2010 Glasgow, UK, September 6-10, 2010.* Berlin: Springer, 2010, s. 124-135. Lecture Notes in Computer Science, Vol. 6273. ISSN 0302-9743. ISBN 3-642-15463-8.

STRODL, Stephan, Petar PETROV a Andreas RAUBER. 2011. *Research on Digital Preservation within projects co-funded by the European Union in the ICT programme* [online]. Vienna: Vienna University of Technology, 2011 [cit. 2012-01-30]. 51 s. Dostupné z: http://cordis.europa.eu/fp7/ict/telearn-digicult/report-research-digital-preservation_en.pdf

ŠVÁSTOVÁ, Pavla. 2010. MEditor – metadatový editor pro digitální knihovnu Kramerius. *Duha* [online]. 25. říjen 2010, č. 3 [cit. 2011-09-10]. ISSN 1804-4255. Dostupné z: http://duha.mzk.cz/clanky/meditor-metadatovy-editor-pro-digitalni-knihovnu-kramerius

TEI CONSORTIUM. 2007. TEI: Text Encoding Initiative. In: *TEI <Text Encoding Initiative>* [online]. TEI Consortium, [2007] [cit. 2011-09-18]. Dostupné z: http://www.tei-c.org/index.xml. Rok vzniku určen dle prvního výskytu v Internet Archive.

TEI CONSORTIUM. 2011a. The TEI Header - TEI P5: Guidelines for Electronic Text Encoding and Interchange. In: *TEI <Text Encoding Initiative>* [online]. TEI Consortium, 3. March 2011 [cit. 2011-09-18]. Dostupné z: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/HD.html

TEI CONSORTIUM. 2011b. The TEI Infrastructure - TEI P5: Guidelines for Electronic Text Encoding and Interchange. In: *TEI <Text Encoding Initiative>* [online]. TEI Consortium, 5. March 2011 [cit. 2011-09-18]. Dostupné z: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ST.html

THURMAN, Alexander C. 2005. Metadata Standards for Archival Control: an Introduction to EAD and EAC. In: *Metadata: a Cataloger's Primer.* Binghamton (NY): Haworth Information Press, 2005, s. 183-201. ISBN 978-0-7890-2801-3. Dostupné také z: http://polaris.gseis.ucla.edu/gleazer/260_readings/Thurman.pdf

Trust. In: *Dictionary.com* [online]. 2011 [cit. 2012-01-18]. Dostupné z: http://dictionary.reference.com/browse/trust

Trusted Digital Repository. In: *TrustedDigitalRepository.eu* [online]. TrustedDigitalRepository.eu, 2010 [cit. 2011-12-12]. Dostupné z: http://www.trusteddigitalrepository.eu/Site/Trusted%20Digital%20Repository.html

UHLÍŘ, Zdeněk. 1999a. K významu a souvislostem přípravy dat pro digitalizaci rukopisů. *Národní knihovna: knihovnická revue.* 1999, roč. 10, č. 3, s. 117-129. ISSN 1214-0678. Dostupné také z: http://full.nkp.cz/nkkr/pdf/9903/9903117.pdf

UHLÍŘ, Zdeněk. 1999b. Projekt MASTER a standardizace v oblasti zpracování rukopisů. *Národní knihovna: knihovnická revue.* 1999, roč. 10, č. 3, s. 109-113. ISSN 0862-7487. Dostupné také z: http://full.nkp.cz/nkkr/pdf/9903/9903109.pdf

UHLÍŘ, Zdeněk. 2002a. Dokončení projektu MASTER. *Národní knihovna: knihovnická revue* [online]. 2002, roč. 13, č. 2, s. 84–101 [cit. 2011-10-02]. ISSN 1214-0678. Dostupné z: http://knihovna.nkp.cz/nkkr0202/0202084.html

UHLÍŘ, Zdeněk. 2002b. *Teorie a metodologie elektronicko-digitálního zpracování rukopisů a hybridní knihovna.* Praha: Národní knihovna ČR, 2002. 324 s. ISBN 80-7050-410-2.

UHLÍŘ, Zdeněk. 2006. *Manuscriptorium v. 1.0: výběr a popis dokumentů* [online]. Verze 1.2. Praha: Národní knihovna ČR, 2006 [cit. 2011-11-25]. 7 s. Dostupné z: http://www.manuscriptorium.com/download/archive/Documentation/manuscriptorium_document_description_CZE.pdf

UHLÍŘ, Zdeněk. 2007. Manuscriptorium na cestě k evropské digitální knihovně. *In: Knihovny současnosti 2007 - sborník* [online]. Brno, 2007 [cit. 2011-09-03], s. 136-144. Dostupné z: http://www.svkos.cz/data/xinha/sdruk/2007-0-136.pdf

UHLÍŘ, Zdeněk. 2010. Evropský projekt ENRICH a jeho význam pro vybudování virtuálního badatelského prostředí. *Knihovna - knihovnická revue* [online]. 2010, roč. 21, č. 1, s. 5-14 [cit. 2011-10-02]. Dostupné z: http://knihovna.nkp.cz/pdf/1001/100105.pdf

VAN DER WERF-DAVELAAR, Titia. 1999. Long-term Preservation of Electronic Publications: the NEDLIB project. *D-Lib Magazine* [online]. September 1999, vol. 5, no. 9 [cit. 2011-09-16]. ISSN 1082-9873. Dostupné z: http://www.dlib.org/dlib/september99/vanderwerf/09vanderwerf.html

VAN GARDEREN, Peter. 2010. Archivematica: using micro-services and open-source software to deliver a comprehensive digital curation solution. In: RAUBER, Andreas, et al. (eds.). *IPRES 2010: Proceedings of the 7th International Conference on Preservation of Digital Objects, Vienna, Austria, September 19-24, 2010.* Vienna: Oesterreichische Computer Gesellschaft, 2010, s. 145-150. ISBN 978-3-85403-262-5.

VERHEUL, Ingeborg. 2006. *Networking for Digital Preservation: Current Practice in 15 National Libraries.* München: Saur, 2006. 269 s. ISBN 3-598-21847-7. Dostupné také z: http://archive.ifla.org/VI/7/pub/IFLAPublication-No119.pdf

VISUAL RESOURCES ASSOCIATION. 2007. Home page. In: *VRA Core* [online]. VRA, 2007 [cit. 2011-09-18]. Dostupné z: http://www.vraweb.org/projects/vracore4/. Rok vzniku odhadnut podle prvního výskytu v Internet Archive.

VOJNAR, Martin. 2006. Standardy digitálních knihoven - nové zkratky. In: *Archivy, knihovny a muzea v digitálním světě 2005.* Praha: Národní technické muzeum, 2006, s. 57-63. Rozpravy Národního technického muzea v Praze, 195. ISBN 80-7037-149-8.

VOJTÁŠEK, Filip. 2000. Dlouhodobá archivace digitálních dokumentů. *Ikaros* [online]. 2000, roč. 4, č. 10 [cit. 2011-09-10]. ISSN 1212-5075. Dostupné z: http://www.ikaros.cz/node/675

VRBICKÝ, Jiří. 2011. *Emailová korespondence ze dne 23. 7. 2011* [online]. [cit. 2011-08-23]

WALKER, Alison. 2006. Preservation. In: BOWMAN, J. H. (ed.). *British Librarianship and Information work 1991–2000.* Hampshire: Ashgate Publishing Limited, 2006. 11 s. (532-548). ISSN 1752-556X. ISBN-10: 0 7546 4779 X. Dostupné také z: http://www.bl.uk/blpac/pdf/bliwa.pdf

WALKER, Alison. 2007. Preservation. In: BOWMAN, J. H. (ed.). *British Librarianship and Information Work 2001-2005.* Hampshire: Ashgate Publishing Limited, 2007. 18 s. (501-518). ISSN 1752-556X. ISBN 978-0-7546-4778-2. Dostupné také z: http://www.bl.uk/blpac/pdf/bliwb.pdf

WATERS, Donald J. 1998. What Are Digital Libraries? *CLIR Issues* [online]. July/August 1998, Nr. 4 [cit. 2011-12-29]. Dostupné z: http://www.clir.org/pubs/issues/issues04.html#dlf

WEBB, Colin. 1999. Preservation Metadata for Digital Collections. In: *National Library of Australia* [online]. Canberra: National Library of Australia, 15. October 1999 [cit. 2011-09-23]. Dostupné z: http://www.nla.gov.au/preserve/pmeta.html

WEBB, Colin. 2003. *Guidelines for the Preservation of Digital Heritage* [online]. Paris: UNESCO, 2003 [cit. 2012-01-01]. 177 s. Dostupné z: http://unesdoc.unesco.org/images/0013/001300/130071e.pdf

WILSON, Andrew. 2007. *Significant Properties Report: InSPECT Work Package 2.2* [online]. V2. London: AHDS, 10/04/2007 [cit. 2012-02-06]. 10 s. Dostupné z: http://www.significantproperties.org.uk/wp22_significant_properties.pdf

WOODLEY, Mary S. 2005. DCMI Glossary. In: *DCMI Home: Dublin Core Metadata Initiative (DCMI)* [online]. Dublin Core Metadata Initiative, 07. 11. 2005 [cit. 2011-10-08]. Dostupné z: http://dublincore.org/documents/usageguide/glossary.shtml

YOTT, Patrick. 2005. Introduction to XML. In: *Metadata: a Cataloger's Primer.* Binghamton (NY): Haworth Information Press, 2005. s. 213- 235. ISBN 978-0-7890-2801-3.

ZARRO, Michael A. a Robert B. ALLEN. 2010. User-Contributed Descriptive Metadata for Libraries and Cultural Institutions. In: LALMAS, Mounia, et al. (Eds.). *Research and Advanced Technology for Digital Libraries. 14ht European Conference, ECDL 2010 Glasgow, UK, September 6-10, 2010.* Berlin: Springer, 2010, s. 46-54. Lecture Notes in Computer Science, Vol. 6273. ISSN 0302-9743. ISBN 3-642-15463-8.

ZENG, Marcia Lei a Jian QIN. 2008. *Metadata.* New York: Neal-Schuman, 2008. 365 s. ISBN 978-1-55570-635-7.

10. Slovník zkratek

Níže jsou uvedeny v textu často používané zkratky. Zkratky uvedené v textu pouze jednou jsou v něm i rozepsány a nejsou v tomto seznamu.

ALTO	Analyzed Layout and Text Object
CASLIN	Czech and Slovak Library Information Network
CCSDS	Consultative Committee for Space Data Systems
CDWA	Categories for the Descriptions of Works of Art
CEDARS	CURL Exemplars in Digital ARchiveS
CIMI	Computer Interchange of Museum Information
CRL	Center for Research Libraries
ČR	Česká republika
DAMS	Digital Asset Management System
DCC	Digital Curation Centre
DIAS	Digital Information Archiving System
DIEPER	DIgitised European PERiodicals
DOBM	Digitized Old Books and Manuscripts – metadatový standard
DPC	Digital Preservation Coalition
DRAMBORA	Digital Repository Audit Method Based on Risk Assessment
DROID	Digital Record Object Identification
DTD	Document Type Definition
EAD	Encoded Archival Description
ECM	Enterprise content management
ENRICH	European Networking Resources and Information Concerning Cultural Heritage
ERPANET	Electronic Resource Preservation and Access Network
EU	Evropská unie
FITS	The File Information Tool Set
FOXML	Fedora Object XML
FP	Framework Programme
GDFR	The Global Digital Formats Registry
IIPC	International Internet Preservation Consortium
iPRES	International Conference on the Preservation of Digital Objects
ISBD	International Standard Bibliographic Description
ISO	International Organization for Standardization
JHOVE	JSTOR/Harvard Object Validation Environment
JIB	Jednotná informační brána
JISC	Joint Information Systems Committee
KB	Koninklijke Bibliotheek – Královská knihovna Nizozemí
KEEP	Keeping Emulation Environments Portable
KNAV	Knihovna Akademie věd Praha
LiWA	Living Web Archives
LMER	Long Term Preservation Metadata for Electronic Resources
LOCKSS	Lots of Copies Keep Stuff Safe

LTP systém	Long-term preservation systém
MARC	MAchine Readable Cataloguing
MASTER	Manuscript Access through Standards for Electronic Records
MC	Master copy – archivní (master) kopie
METS	Metadata Encoding and Transmission Standard
MIX	Metadata for Images in XML
MMSB	Memoriae Mundi Series Bohemica
MODS	Metadata Object Description Scheme
MZK	Moravská zemská knihovna Brno
NDIIPP	National Digital Information Infrastructure and Preservation Program
NDK	Národní digitální knihovna – projekt
NEDLIB	Networked Europian Deposit Library
Nestor	Network of Expertise in Long-Term Storage of Digital Resources
NISO	National Information Standards Organization
NK ČR	Národní knihovna České republiky
NK NZ	Národní knihovna Nového Zélandu
NRG	National Representatives Group
NZME	New Zealand Metadata Extractor
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OAIS	Open Archival Information System
OCR	Optical Character Recognition
PLANETS	Preservation and Long-term Access through NETworked Services
PLATTER	Planning Tool for Trusted Electronic Repositories
PREMIS	PREservation Metadata Implementation Strategies
RLG	Research Libraries Group
SDB	Safety Deposit Box – LTP systém od firmy Tessella, UK
SGML	Standard Generalized Markup Language
SICI	Serial Item and Contribution Identifier
SRU/SRW	Search and Retrieve via URL/Search and Retrieve via Web
TEI	Text Encoding Initiative
TOTEM	Trustworthy Online Technical Environment Metadata
TRAC	Trustworthy Repositories Audit & Certification
UC	User copy – uživatelská kopie
UDFR	Unified Digital Formats Registry
UUID	Universally Unique IDentifier
VaV	Výzkum a vývoj – programové financování z Ministerstva kultury ČR
VISK	Veřejné informační služby knihoven
VRA	Visual Resoucrce Association
WARC	Web ARChive – kontejnerový standard pro uložení dat ze sklizně webu
WG	Work Group – pracovní skupina
WP	Work Package – část projektu
XML	Extensible Markup Language
XSD	XML Schema Definition
XSLT	eXtensible Stylesheet Language Transformations

11. Příloha – Návrhy metadatových profilů pro digitalizaci v projektu NDK

Navržené profily pro digitalizaci monografií a periodik vycházejí ze stejného základu a některé jejich části se proto překrývají. I přesto jsou oba profily uvedeny jako samostatný celek, podobně, jak byly vytvořeny pro projekt NDK a také publikovány – viz [HUTAŘ, 2011a; HUTAŘ, 2011b]. Jejich případné další využití pro zájemce mimo projekt NDK tak bude jednodušší, než kdyby byly oba profily uměle spojeny do jednoho celku.
Každý profil tak obsahuje úvodní obecné části, seznam elementů jednotlivých použitých schémat, jejich popis (vysvětlení, povinnost použití, opakování, možnosti přebírání metadat, syntax zápisu jejich hodnot, atributy, doporučení o použití kontrolovaných slovníků). Profil na několika místech pouze konstatuje, že bude nutné kontrolované slovníky vytvořit, někde uvádí i návrh jednotlivých položek. Některé kontrolované slovníky jsou pro každou instituci mnohdy odlišné, odvíjející se od lokálních zvyklostí a skutečností. Nutné bude také dopracovat pravidla plnění elementů v jednotlivých institucích nebo projektech.
Pro tvorbu metadat podle předkládaného profilu je uzpůsoben v ČR SW Sirius od firmy ELSYST, která jej využívá pro projekt ANL+, který specifikaci metadatového profilu pro NDK převzal. Ze světových systémů by tuto specifikaci měl být schopen vytvořit každý SW, který podporuje kompletní sady elementů v profilu užívaných schémat, tedy METS, MIX, MODS, PREMIS a ALTO XML, včetně Dublin Core. Jmenovat lze například SW DocWorks od firmy CCS.

Oba profily metadat byly ve finálních fázích jejich tvorby konzultovány s kolegy z NK ČR (Marek Melichar, Bedřich Vychodil, Jiří Polišenský, Ivan Ljubka) a z Moravské zemské knihovny v Brně (Pavla Švástová). Díky za cenné připomínky patří také pracovníkům firmy Elsyst, kteří přinesli řadu podnětů při reálném využití profilu pro digitalizaci periodik v projektu ANL+ na sklonku roku 2011.
Oba profily jsou ve vývojových verzích, profil pro monografie ve verzi 0.3, profil pro periodika ve verzi 1.2.

11.1 Aplikační metadatový profil pro digitalizaci periodik

Definice metadatového profilu pro digitalizaci periodik v projektu NDK

VERZE 1.2 - 12.2.2012

1. Východiska

- uživatelské kopie = UC
- archivní kopie = MC
- původní sken = PS – obrazový soubor vzniklý při digitalizaci, který se po zpracování (ořez, narovnání apod.); maže se ještě v procesu digitalizace a dále se neukládá
- u všech metadatových standardů budou použity verze aktuální v době implementace projektu NDK, nebo verze předchozí v případě, že nová verze je nová min. 3 měsíce
- základní intelektuální entita ve workflow digitalizace a následně i v LTP systému = číslo periodika
- PSP balíček – *producer submission package*
    - balíček dat a metadat, který přichází od producenta dat (z workflow digitalizace)
    - PSP balíček bude obsahovat kompletní intelektuální entitu tj. číslo periodika
    - z workflow digitalizace lze poslat více PSP balíčků v balíku např. [.tar] apod.
    - pokud má vícesvazkové dílo v katalogu knihovny bibliografický záznam pro každý svazek, vznikne pro každý svazek PSP balíček a každý svazek bude brán jako jedna intelektuální entita; to samé platí i pro případ, že vícesvazkové dílo má pouze jeden záznam
- základní bibliografická metadata budou stahována do workflow digitalizace z knihovních katalogů
- jako výchozí SW pro vytváření souborů JPEG 2000 se bude používat Kakadu
- úpravy obrazu, které vedou ke změně rozměrů obrazu, rozlišení apod., se musí dělat před tím, než se udělá OCR, tj. budou se dělat na TIFF souborech
- veškerá metadata musí pro zápis používat kódování UTF-8

2. Výstupy digitalizace

1. archivní kopie (1 MC pro každou stránku)
2. uživatelské kopie (1 UC pro každou vzniklou MC, tedy stránku)
3. OCR – ALTO XML soubor pro každou stránku
4. OCR TXT soubor – pro možnost stáhnout si jen text dokumentu (tam kde kvalita OCR je odpovídající), vyhledávání/indexace
5. metadata pro MC
    - bibliografická metadata – MODS a DC
    - strukturální metadata – METS
    - technická metadata – MIX, PREMIS
    - administrativní metadata – PREMIS, METS
6. kontrolní metadatové soubory (s kontrolními součty a údaji o vzniku dat apod.)

Pozn.
Záznam METS nebude obsahovat žádná metadata pro uživatelské kopie. METS neobsahuje popisná, ani technická metadata pro UC. Obrazové soubory UC nejsou ani součástí strukturální mapy <structMap> ani <fileSec>. Součástí PSP balíčku budou jen obrazy UC ve složce [userCopy].

Důvodem je to, že metadata pro UC budou vytvářena na vstupu do Krameria4 ve standardu FOXML (Fedora Object XML). Budou se vyrábět z METS záznamu pro MC, jehož specifikace je níže.

3. Granularita metadatového záznamu

Periodika

- základní intelektuální entitou periodik je 1 číslo
- každé číslo periodika má svůj vlastní metadatový záznam (METS), který obsahuje údaje o nadřazených entitách čísla, jako jsou ročník, titul periodika; tj. je pro uživatele i pro systém možné spojit jednotlivá čísla do ročníků a titulů

4. Identifikátory

Do workflow digitalizace budou přicházet bibliografická metadata, která již mohou obsahovat následující identifikátory vrchních úrovní intelektuálních entit (úroveň titulu):

- ISBN – pro titul monografie (jednosvazkové); nebo pro soubor monografií, které mají pouze jeden souborný záznam, ISBN není přiděleno vždy;
- ISSN – pro titul periodika, ISSN není přiděleno vždy (chybí např. u starých titulů z 19. století);
- čČNB – identifikátor entity tak jak odpovídá katalogizačnímu záznamu, tj. každá entita se záznamem v katalogu NK/MZK má tento identifikátor.

Pokud není dostupný ani jeden z výše uvedených, lze použít čárový kód dokumentu, systémové číslo, signaturu, nebo systémové číslo kombinované s polem 001 MARC záznamu apod. Jednou z možností je také využití jednoznačného čísla UUID.

Nižší úrovně intelektuálních entit by měly mít také své identifikátory, ideálně URN:NBN (pouze pro digitální dokumenty), které bude přidělováno během digitalizace[290]. Přidělováno bude logickým úrovním (entitám). U periodik tedy číslu, vnitřní části, příloze. Syntax URN:NBN musí odpovídat specifikaci identifikátoru URN:NBN pro resolver NK ČR (např. urn:nbn:cz:ndk-123456 pro výstupy z projektu NDK).

5. Struktura PSP balíčku

Níže je podoba struktury balení dat a metadat v jednom PSP balíčku na výstupu z workflow digitalizace. PSP balíček = 1 složka pro 1 číslo periodika.

[290] Buď přímo v SW pro workflow digitalizace, nebo za pomoci aplikace Resolver URN:NBN.

složka	obsahuje >	obsahuje >	obsahuje>	
svazek monografie /číslo periodika				
	info.xml	údaje o vzniku balíku		
	složka [masterCopy]	obrazy JPEG2000 lossless		
	složka [userCopy]	obrazy JPEG2000 lossy		
	složka [ALTO]	soubory ALTO XML		
	složka [TXT]	soubory OCR.TXT		
	složka [amdSec]	AMD_METS. xml soubor pro každou stránku obsahuje>	amdSec	**techMD** = PREMISobject pro MC, původní TIFF, ALTO XML) + MIX pro MC, původní TIFF)
				digiprovMD = PREMISevent + PREMISagent
			fileSec	odkazuje na MC, ALTO XML, OCR TXT soubor popisované 1 stránky
			StructMap	pouze fyzická - pro soubory popisované stránky (MC a ALTO XML, OCR TXT)
	Hlavni_METS.xml	dmdSec	MODS a DC pro jednotlivé úrovně dokumentu	
		fileSec	obsahuje linky na MC, ALTO XML, OCR TXT a technická metadata ve složce [amdSec]	
		structMap (včetně ALTO odkazů)	logická a fyzická pro MC, ALTO XML areas, OCR TXT a AMD_METS.xml	
	MD5	kontrolní součty všech souborů v PSP balíku		

Jedná se o variantu, kdy technická a administrativní metadata nejsou obsažena v hlavním METS záznamu, ale pro každou stránku v jiném dalším METS záznamu (AMD_METS.xml). Důvodem je to, že pokud by bylo vše v hlavním METSu, byl by neúměrně dlouhý. Takto je z hlavního záznamu vedlejší METS nalinkován.

Hlavní složka PSP balíčku obsahuje následující složky a soubory:

soubor info.xml

Velmi krátce tu budou zaznamenány údaje o vzniku celého PSP balíčku – kdo a kdy ho vytvořil, jakou měl velikost, odkud kam byl nakopírován apod. Obsahovat by také měl informaci o stavu zpracování balíčku. Zaznamenány by také měly být údaje o obsahu PSP balíčku – počet a názvy souborů apod. Soubor info.xml by také mohl být vedle hlavního PSP balíčku. Údaje a struktura info.xml souboru:

1. vznik balíčku – datum dle ISO8601 na úroveň vteřin
2. ID balíčku – použít identifikátor čísla periodika (URN:NBN) – viz názvová konvence v kapitole 6
3. ID titulu - čČNB, ISBN nebo ISSN
4. údaje o větším celku, do kterého balíček patří - např. digitalizace pro ANL
5. název instituce, která je zadavatelem digitalizace

6. tvůrce balíčku – kód instituce (firmy), která balíček vytvořila
7. velikost balíčku – v kB
8. z jakého serveru bylo nahráno – URL
9. obsah balíčku
    - názvy souborů včetně directory path a koncovky (MIME type)
    - počet souborů v balíčku celkem
10. stav zpracování – možné hodnoty
    - hotovo
    - opraveno
        - added OCR
        - added titles
        - added logical parts (issues, years)
        - updated xml (Mods, DC, identifikátory),
11. poznámka – např. o tom, že balíček neobsahuje OCR apod.

Příklad balíčku, který obsahuje 2 soubory, jeden v kořenu složky a druhý ve složce:

1. CREATED=2009-11-10T12:37:46
2. PACKAGEID=ANL_123456
3. TITLEID=ISSN1234-1236
4. COLLECTION=ANL
5. INSTITUTION=NKP
6. CREATOR=NazevFirmy
7. PACKAGESIZE=36000155kb
8. SOURCELOCATION= server123.firma.cz/baliky_hotovo/01/2011/12/000025456
9. ITEMLIST=scan01.jp2
   ITEMLIST=slozka/hotovo/27.9.2011/scan02.jp2
   ITEMTOTAL=2
10. STATUS=hotovo
11. NOTE=noOCR

složka [masterCopy]
Složka s archivními kopiemi, obsahuje soubory JPEG 2000 v neztrátové kompresi, 1 soubor = 1 stránka, tj. obsahuje všechny naskenované stránky čísla periodika.

složka [userCopy]
Složka s uživatelskými kopiemi, pro každou naskenovanou stránku čísla periodika obsahuje jeden JPEG 2000 soubor se ztrátovou kompresí.

složka [ALTO]
Obsahuje ke každé stránce 1 ALTO XML soubor, tj. tolik ALTO XML souborů kolik je stránek čísla periodika.

složka [TXT]
Obsahuje ke každé stránce 1 OCR soubor jako čistý text. Tj. tolik OCR.TXT souborů, kolik je stránek čísla periodika.

složka [amdSec]

Složka s technickými metadaty – **obsahuje pro každou naskenovanou stránku čísla časopisu 1 METS soubor/záznam (AMD_METS.xml)**. Záměrně nejsou tato metadata v hlavním METS záznamu (hlavni_METS.xml), musí z něj být ovšem nalinkována (z části fileSec). Každý METS záznam AMD_METS.xml obsahuje následující části METS standardu:

- amdSec – administrativní metadata – obsahuje část:
  - technických metadat (techMD), která ve standardu PREMIS Object popisuje vlastnosti archivních kopií, ALTO XML, původního TIFF souboru, ze kterého vznikly archivní kopie. Dále je přítomen záznam technických metadat v MIX standardu pro archivní kopie a pro původní TIFF.
  - metadat o provenienci digitálních objektů (digiprovMD) – v této části je využit standard PREMIS Event a PREMIS Agent.
  - fileSec- sekce s odkazy na soubory. V případě tohoto METS záznamu pro jednu stránku, který vzniká primárně k zachycení technických a administrativních metadat, bude odkazovat na soubory, které jsou s tou konkrétní stránkou spojeny, tj. archivní kopie, ALTO XML a OCR TXT. Jde o povinnou sekci METS záznamu.
  - structMap – pouze fyzická strukturální mapa, povinná část METS záznamu. Bude ukazovat strukturu souborů k dané stránce, tj. opět archivní kopie, ALTO XML a OCR TXT. Pro další mapování do LTP systému nebude potřeba.

soubor Hlavni_METS.xml

Další částí PSP balíčku je hlavní METS dokument. Hlavní METS záznam obsahuje:

- dmdSec – bibliografická metadata k číslu periodika včetně popisu nadřazených entit (např. ročník, titul) nebo naopak částí (např. kapitola). Základ bude z katalogu, případný další popis částí vznikne v procesu digitalizace. Hlavní standardem metadat je MODS, nutná pro LTP je i přítomnost zkráceného záznam v Dublin Core.
- fileSec – hlavní část s linky na všechny digitální objekty (archivní kopie, ALTO XML a OCR TXT), které se váží k jednomu číslu periodika. Obsahuje také linky na administrativní metadata AMD_METS.xml do složky [amdSec].
- structMap – strukturální mapa pro celý dokument, tj. pro jedno číslo periodika. Obsahuje:
  - logickou část – vyjadřuje logickou strukturu čísla periodika s odkazy na ALTO XML;
  - fyzickou část obsahující informace o všech reprezentacích konkrétní stránky (archivní kopie, ALTO XML, OCR TXT a AMD_METS.xml);
  - mapování na ALTO XML areas.

soubor MD5

Poslední částí PSP balíčku je soubor s kontrolními součty pro všechny soubory balíčku (kromě info.xml a .md5 souboru samotného). Soubor .md5 je jeden pro 1 celý balíček PSP (balíček s číslem periodika). Tento soubor .md5 obsahuje kontrolní součet pro každý soubor obsažený v PSP balíčku. Z tohoto důvodu nejsou samostatné kontrolní součty součástí podsložek balíčku. Kontrolní součty jsou také samozřejmě v technických metadatech.

6. Názvová konvence složek a souborů

pojmenování PSP balíčku

- každý PSP balíček přicházející z digitalizace by měl obsahovat pouze jedinou intelektuální entitu (číslo periodika). **Pak musí název balíčku vycházet z identifikátoru této entity, např. URN:NBN, číslo čárového kódu použitého na fyzické jednotce apod.**
- **každé číslo periodika musí mít svůj jednoznačný identifikátor, tím pádem pak každý PSP balíček a každý soubor v něm má vlastní jednoznačný identifikátor**

pojmenování složek

- viz návrh struktur PSP balíčku (kapitola 5)

pojmenování souborů

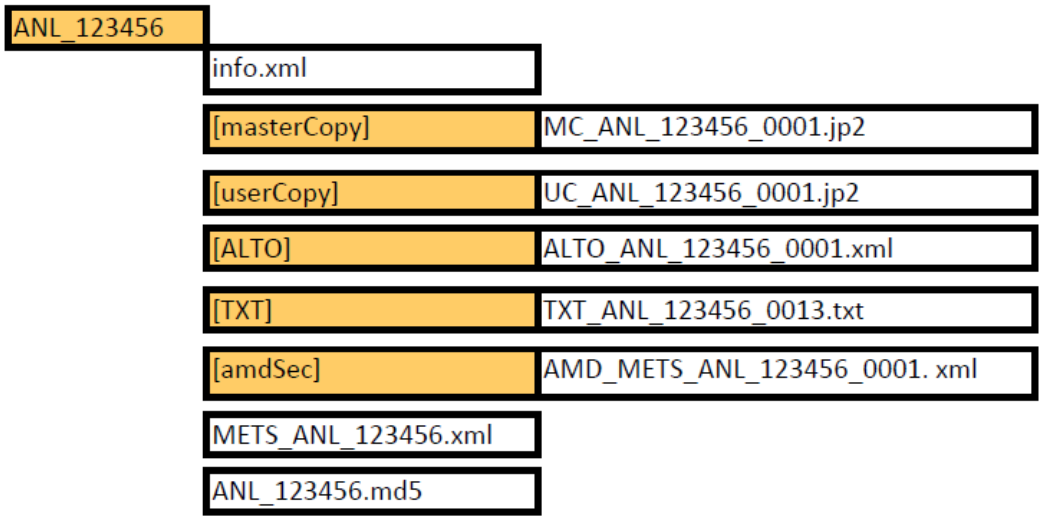
- názvy jakýchkoliv souborů náležejících k jedné základní entitě (svazek nebo číslo) musí být založeny na jednom typu identifikátoru
- pro číslo periodika by takovým identifikátorem mohlo být URN:NBN, čČNB, ISBN nebo ISSN titulu + další upřesnění (číslo výtisku apod.)
- podobně využitelným identifikátorem by mohlo být generované číslo UUID, které by se generovalo pro každý soubor; tím by se ovšem ztratila vazba (i vizuální) na vrchní úroveň titulu i vazba na související soubory (stránka v jp2 a k ní náležející soubor ALTO XML apod.)

S využitím URN:NBN to může vypadat následovně (použit příklad pojmenování pro projekt ANL+ digitalizace periodik):

typ souboru	název souboru	vysvětlení
PSP balíček (číslo, svazek)	ANL_123456	název celé složky PSP balíčku, u základních intelektuálních entit bude v názvu využito vždy URN:NBN
archivní kopie	MC_ANL_123456_0013.jp2	archivní JPEG 2000 stránky 13 čísla periodika s urn:nbn:cz:anl-123456
uživatelská kopie	UC_ANL_123456_0013.jp2	uživatelská kopie ve formátu JPEG 2000 stránky 13 čísla periodika s urn:nbn:cz:anl-123456
ALTO XML	ALTO_ANL_123456_0013.xml	ALTO soubor náležející ke 13té stránce z čísla periodika s urn:nbn:cz:anl-123456
OCR TXT	TXT_ANL_123456_0013.txt	TXT soubor s OCR náležející ke 13té stránce z čísla periodika s urn:nbn:cz:anl-123456
info.xml	INFO_ANL_123456.xml	informační XML k celému PSP balíčku čísla periodika
MD5	ANL_123456.md5	soubor s kontrolními součty k celému PSP balíčku čísla periodika

Hlavni_METS.xml	METS_ANL_123456.xml	hlavní METS záznam k celému číslu periodika s urn:nbn:cz:anl-123456
AMD_METS.xml	AMD_METS_ANL_123456_0013.xml	METS záznam s technickými metadaty pro stránku 13 z čísla periodika s urn:nbn:cz:anl-123456

Složka jednoho balíčku PSP, který obsahuje jen jeden obrazový soubor k první stránce čísla periodika, pak může vypadat následovně (příklad balíčku z digitalizace v projektu ANL+):



7. Transportní balík pro jeden nebo více PSP balíčků

Pokud bude jeden PSP balík obsahující 1 základní intelektuální entitu (číslo periodika) přemísťován např. jako tar soubor, měl by název souboru tar odpovídat názvu PSP balíčku (tedy vycházet z použitého identifikátoru pro číslo).
Výstupem workflow digitalizace ale může také být balík (např. tar), který obsahuje více PSP balíčků. Toto sdružování bude omezeno jen kapacitou HW. Takovýto sdružený balík by měl být pojmenován na základě již užívaného identifikátoru.

- v případě, že balík obsahuje čísla periodika, měl by název balíku vycházet z čČNB nebo z ISSN
- v případě, že balík obsahuje svazky vícesvazkového díla, měl by název balíku vycházet z čČNB nebo ISBN
- typ identifikátoru musí být vyjádřen v názvu souboru – např. ISSN_1234-5678.tar nebo CCNB_12345678910.tar apod.
- lze počítat s tím, že bude docházet k tomu, že sdružený balík nebude obsahovat např. všechny čísla určitého titulu periodika – tato skutečnost musí být patrná z názvu balíku (např. ISSN_1234-5678_YYYY; kde YYYY může být pořadové číslo, datum, doba vzniku jednoho z více balíků obsahujících čísla určitého titulu s identifikátorem ISSN 1234-5678)

Transportní balík by měl obsahovat následující části:

- balíčky PSP (svazků nebo čísel);
- informační soubor, který odpovídá specifikaci info.xml;

- kontrolní součty všech PSP balíčků;
- seznam balíčků v transportním balíku.

8. Metadata

- veškerá metadata budou „zabalena“ pomocí kontejnerového standardu METS
- standard METS bude v aktuální verzi v době implementace nebo verzi předchozí (prosinec 2010 verze 1.9 - http://www.loc.gov/standards/mets/mets-schemadocs.html)
- veškerá metadata ve všech standardech metadat musí být zapsána pomocí XML za použití kódování UTF-8
- **vložení metadatových schémat do kontejneru METS bude vždy formou <mdWrap>, tj. ne odkazováním z METS záznamu ven**

8.1 Vysvětlivky k tabulkám

Obsah pole „Použití pro“
použití jednotlivých elementů pro popis MC, PS (původní sken), XML (ALTO)

Pole „Popis“ obsahuje:
- vysvětlení a příklad
- doporučené plnění tam, kde je to možné uvést
- vysvětlení a doporučené hodnoty atributů
- opakovatelnost výskytu elementu (např. pro standard PREMIS – dle XSD)
    - 0-1 element je nepovinný, neopakovatelný
    - 0-n element je nepovinný, opakovatelný
    - 1-n element je povinný a opakovatelný
    - 1 element je povinný a neopakovatelný

Význam pole „povinnost“
Pole „povinnost“ uvádí, zda je plnění jednotlivých elementů povinné, doporučené nebo volitelné. Může nabývat následujících hodnot:

1. M mandatory (povinně plnit – element je součástí každého záznamu)
2. MA mandatory if available (povinně plnit pokud je to možné, pokud lze apod.)
3. R recommended (plnění hodnot elementu je doporučeno, není ovšem povinné)
4. RA recommended if available (doporučeno pokud lze plnit)
5. O optional (plnění hodnot elementu je zcela dle konkrétních potřeb)

- v případě tabulky obsahující dvě schémata (např. MODS a Dublin Core) platí povinnost pro elementy obou schémat stejně
- pokud je rodičovský element např. doporučený a dceřiný element povinný, znamená to, že dceřiný element je povinný pouze tehdy, pokud je použit element rodičovský
- oranžová barva v tabulkách označuje elementy, které mají povinný výskyt

8.2 Kořenový element hlavního METS záznamu

element	atributy	popis	Povinnost
<mets>	LABEL TYPE	kořenový element METS záznamu --- LABEL – název titulu periodika, včetně čísla a data vydání čísla, např. Mladá fronta no. 5 29.06.1979 TYPE – hodnota vždy „Periodical“	M

Kořenový element hlavního METS záznamu k jednotlivému číslu periodika musí obsahovat linky na specifikace jednotlivých použitých metadatových schémat (METS, MODS, Dublin Core).

8.3 METS hlavička <metsHdr>

Dokumentuje vznik a úpravy METS záznamu.

element	atributy	popis	Povinnost
<metsHdr>	LASTMODDATE CREATEDATE	hlavička METS záznamu --- LASTMODDATE – datum poslední úpravy záznamu, musí být ve tvaru ISO 8601 (na úroveň vteřin) CREATEDATE – datum vytvoření záznamu, musí být ve tvaru ISO 8601 (na úroveň vteřin)	M
<agent>	ROLE TYPE	údaje o tvůrci záznamu METS --- ROLE – hodnota „CREATOR“ TYPE – hodnota „ORGANIZATION“	M
<name>		jméno jednotlivce nebo organizace; --- tvůrce záznamu, buď dodavatel (firma XY) nebo v případě tvorby záznamu v knihovně bude využita sigla knihoven, tj. pro NK ČR hodnota „ABA001“	M

8.4 METS část <dmdSec> - Bibliografická metadata periodik (MODS a Dublin Core)

- na samotný bibliografický popis bude použit standard MODS, aktuální verze v době implementace, nebo verze předchozí (prosinec 2010 verze 3.4 viz http://www.loc.gov/standards/mods/) a nekvalifikovaný Dublin Core (DC) (http://dublincore.org/documents/dcmi-terms/)
- DC je primárně určeno na poskytnutí dat přes OAI-PMH, bude odpovídat OAI XSD (viz http://www.openarchives.org/OAI/2.0/oai_dc.xsd)
- DC bude uloženo v METS apod. stejným způsobem jako standard MODS – viz struktura PSP balíčku výše
- pro vytvoření DC z MODS standardu může být použito oficiální mapování Kongresové knihovny – viz http://www.loc.gov/standards/mods/mods-conversions.html
- DC a MODS bude vložen v METS části dmdSec – viz struktura PSP balíčku v kapitole 5
- základním zdrojem pro popisná metadata je katalog NK ČR a MZK
- u digitalizovaných dokumentů je bibliografický popis vytvářen primárně z pohledu popisu fyzické předlohy, nejde o popis elektronického dokumentu

Základní intelektuální entitou pro popis je číslo periodika, tj. v jednom METS záznamu, který bude obsahovat metadata a strukturu jednoho čísla periodika, budou MODS záznamy k tomuto číslu. Metadata budou popisovat následující entity:

- ***titul (Title)***
- ***číslo (Issue)***
- ***vnitřní část (InternalPart)*** – typy článek (Article) a obraz (Picture)
- ***příloha (Supplement)***

- ***ad titul (Title)*** – MODS záznam bude obsahovat i číslo ročníku
- ***ad číslo (Issue)*** – typy čísla jsou v elementu <genre> za použití atributu type
- ***ad vnitřní část (InternalPart)*** – typy vnitřní části článek a obraz by měly pokrýt veškerou variabilitu možností, které mohou texty a obrázky na tištěné stránce mít; bližší určení typů článku (novinky, zprávy, reklama apod.) a obrazu (fotografie, tabulka, ilustrace, graf apod.) bude možné vyjádřit pomocí atributů a výrazů kontrolovaného slovníku v elementu <genre>
- ***ad příloha (Supplement)*** – přílohou se rozumí volně vložená entita do jednotlivého čísla, např. mapa, obsah celého ročníku, CD/DVD apod. Rozlišujeme 3 druhy příloh periodik:
    1. příloha, která **se neskenuje**, ale chceme o ní vytvořit bibliografický záznam; dát najevo čtenáři, že existuje (např. CD/DVD, pohlednice, plakát apod.).
        - digitální podoba přílohy (pokud existuje) není součástí balíčku PSP čísla (Issue)
        - popis lze udělat v rámci popisu přílohy (Supplement) v MODS – viz specifikace níže
        - taková příloha není součástí logické strukturální mapy standardu METS
    2. příloha podobného typu, tvaru a velikosti jako je popisované číslo periodika, která se spolu s číslem **skenuje.**

- digitální podoba přílohy je, spolu s číslem (Issue), součástí PSP balíčku čísla (Issue) a je součástí hlavního METS záznamu
- popis lze udělat v rámci popisu přílohy (Supplement) v MODS – viz specifikace níže
- taková příloha může mít vnitřní části (InternalPart) stejně jako číslo (Issue) a jejich text je součástí ALTO XML, které je společné pro číslo (Issue) i přílohu (Supplement)
- **taková příloha je součástí logické strukturální mapy standardu METS**
- **taková příloha je součástí fyzické strukturální mapy standardu METS (linky mezi jednotlivými soubory reprezentujícími stránky a popisnými metadaty)**

3. příloha odlišného typu, tvaru a velikosti než je popisované číslo periodika, která **se skenuje a popisuje zvlášť** na čísle nezávisle.
   - taková příloha se zpracovává z pohledu katalogizace jako „nezávislé“ periodikum, z pohledu digitalizace pak jako „nezávislý“ časopis
   - může se jednat o přílohy časopiseckého typu vycházející u různých deníků (Pátek u Lidových Novin, čtvrteční příloha MF Dnes apod.)
   - k těmto přílohám vznikají metadata podobně jako pro jednotlivá čísla deníků nebo klasické časopisy, ovšem na původním čísle, ke kterému příloha patřila, nezávisle; tj. pro „původní“ číslo, u kterého byla příloha, vznikne 1 popis (PSP balíček s jedním hlavním METS záznamem a ALTO XML souborem) a pro přílohu je vytvořen další 1 popis (a PSP balíček s METS záznamem), jako by šlo o běžný samostatný časopis
   - příloha se pak popisuje jako číslo (Issue)

- jednotlivé MODS záznamy pro části (titul, číslo, vnitřní část a příloha) nejsou samopopisné, tj. neobsahují vždy údaje o vrchních entitách (článek neobsahuje informace o titulu apod.)
- pro každou entitu/úroveň záznamu vznikne jeden MODS záznam s vlastním ID, které bude označovat i typ části (např. článek, ilustrace apod.); v případě opakování částí se bude opakovat odpovídající počet MODS záznamů
- každý MODS záznam bude uložen ve vlastní METS části <dmdSec> pomocí <mdWrap>
- u úrovní kde je to potřeba (vnitřní část, příloha apod.) se budou opakovat <dmdSec> části tolikrát, kolik je konkrétních částí
  - tj. v METS záznamu vznikne 1 část <dmdSec> pro bibliografický záznam titulu periodika, 1 <dmdSec> část pro bibliografický záznam čísla periodika, několik <dmdSec> části pro vnitřní části (pro všechny články i obrázky) a odpovídající počet <dmdSec> částí pro přílohy, dle počtu příloh
  - bibliografický popis obrazů bude velmi minimalistický
- v katalozích NK ČRa MZK existuje záznam pouze pro titul periodika, neexistují samostatné záznamy pro čísla, ročníky apod. – tj. vnitřní členění a popis musí vzniknout v digitalizaci, popis titulu periodika musí být stažen z katalogu do workflow digitalizace
- stránka se nebude popisovat, její logické i fyzické číslování i typ stránky je obsaženo ve struktuře METS dokumentu (část structMap)
  - typ stránky (Advertisement, Blank, Index aj.) budou odpovídat přesně seznamu typů z DTD periodika – viz http://digit.nkp.cz/DigitizedPeriodicals/DTD/2.10/Periodical.xsd

- **všechny top elementy MODS standardu jsou opakovatelné, kromě <recordInfo>**
- **všechny elementy Dublin Core jsou opakovatelné**

Každá část <dmdSec> musí mít ID (identifikátor) a vnořený element <mdWrap> s atributy MDTYPE, MIMETYPE.

element	atributy	popis	povinnost
<dmdSec>	ID	identifikátor <dmdSec> části METS záznamu --- ID: pro <dmdSec> s popisem titulu periodika hodnota „MODSMD_TITLE“ pro záznam v MODS nebo „DCMD_TITLE“ pro záznam v Dublin Core pro <dmdSec> s popisem čísla periodika hodnota „MODSMD_ISSUE“ a „DCMD_ISSUE“ pro <dmdSec> s popisem vnitřní části periodika hodnota dle typů vnitřní části (článek, obraz) - hodnoty „MODSMD_ART“ a „DCMD_ART“ pro článek a hodnoty „MODSMD_PICT“ a „DCMD_PICT“ pro obraz pro <dmdSec> s popisem přílohy periodika hodnota „MODSMD_SUPPL“ a „DCMD_SUPPL“	M
<mdWrap>	MDTYPE MIMETYPE	element obsahující vložené záznamy MODS --- MDTYPE – hodnota „MODS“ pro záznamy v MODS, hodnota „DC“ pro záznam v Dublin Core MIMETYPE – hodnota „text/xml“	M

8.4.1 Pole MODS a Dublin Core pro jednotlivé části periodika

8.4.1.1 Pole MODS a DC pro titul periodika

Element MODS	Atributy	Popis	povinnost	Element DC
<titleInfo>	ID	název titulu periodika pro plnění použít katalogizační záznam --- ID musí vyjadřovat název úrovně,	M	

		tj. např. „MODS_TITLE“		
<title>		názvová informace – název periodika hodnoty převzít z katalogu	M	<dc.title>
<subTitle>		podnázev periodika	MA	<dc.title>
<partNumber>		číslo části, např. určité řady/edice (část 1, řada B), k použití u ročenek apod.	R	<dc:description>
<partName>		jméno edice nebo speciální ediční řady, např. Hygiena. k použití u ročenek a specializovaných periodik	R	<dc:description>
<typeOfResource>		popis charakteristiky typu nebo obsahu zdroje jedna z hodnot: - text - cartographic - notated music - sound recording-musical - sound recording-nonmusical - sound recording - still image - moving image - three dimensional object - software, multimedia - mixed material pro periodika a monografie hodnota text; mělo by se vyčítat z MARC21 katalogizačního záznamu z pozice 06 návěští	R	<dc:type>
<genre>		bližší údaje o typu dokumentu hodnota: title	M	<dc:type>
<originInfo>		informace o původu předlohy Poznámka: Jeden nebo více výskytů elementů se předpokládá pro vydavatele, další výskyt v případě nutnosti popsat tiskaře. Pokud je nutno	M	

		vyjádřit tiskaře (pole 260 podpole „f“ a „e“ a „g“ v MARC21), je nutno element <originInfo> opakovat s atributem transliteration=“printer“ a elementy <place>, <publisher>, <dateCreated>, které budou obsahovat údaje o tiskaři. Pokud bylo za dobu vydávání více vydavatelů, nutno vzít z katalogizačního záznamu pole 260 indikátor 02 a údaje o vydavatelích opakovat.		
<place>		údaje o místě spojeném s vydáním, výrobou nebo původem popisovaného dokumentu	MA	<dc:coverage>
<placeTerm>	type	konkrétní určení místa, např. Praha odpovídá hodnotě z katalogizačního záznamu, pole 260, podpole „a“ ---------- type – bude vždy text	MA	<dc:coverage>
<publisher>		jméno entity, která dokument vydala, vytiskla nebo jinak vyprodukovala odpovídá poli 260 podpoli „b“ katalogizačního záznamu v MARC21; v případě, že existovalo více vydavatelů, jsou uvedeni v poznámce v poli 500 a měli by se objevit v elementu top elementu <note>	MA	<dc:publisher>
<dateIssued>		datum vydání předlohy, nutno zaznamenat v případě titulu roky v nichž časopis vycházel (např. 1900-1939), přebírat ve formě, jak je zapsáno v hodnotě pole v katalogu odpovídá hodnotě z katalogizačního záznamu, pole 260, podpole „c“	M	<dc:date>

<dateCreated>		datum vytvoření předlohy bude použito pouze při popisu tiskaře, viz poznámka u elementu <originInfo> odpovídá hodnotě z katalogizačního záznamu, pole 260, podpole „g“	R	
<issuance>		údaje o vydávání hodnota continuing odpovídá hodnotě uvedené návěští MARC21 na pozici 07	M	
<frequency>		údaje o pravidelnosti vydávání odpovídá údaji MARC21 v poli 310 nebo pozici 18 v poli 008	R	
<language>		údaje o jazyce dokumentu	M	
<languageTerm>	type authority	přesné určení jazyka – kódem nutno použít kontrolovaný slovník ISO 639-2, http://www.loc.gov/standards/iso639-2/php/code_list.php --- type: použít hodnotu code authority: použít hodnotu „iso639-2b“	M	<dc:language>
<physicalDescription>		obsahuje údaje o fyzickém popisu zdroje/předlohy	M	
<form>	authority	údaje o fyzické podobě dokumentu, např. print, electronic apod. pro periodika hodnota print odpovídá hodnotám pozice 23 a 29 v poli 008 MARC21 --- authority: hodnota „marcform“	M	<dc:format>
<extent>		údaje o rozsahu (stran, svazků nebo rozměrů); použití spíše u ročenek apod. odpovídá hodnotám v poli 300 podpolích „a“ a „c“ MARC21, pokud jsou vyplněna obě pole,	RA	<dc:format>

		bude se element <extent> opakovat		
<note>		poznámka o fyzickém stavu dokumentu; pro každou poznámku je nutno vytvořit nový <note> element	RA	
<abstract>		shrnutí obsahu periodika jako celku odpovídá poli 520 MARC21	R	<dc:description>
<note>		obecná poznámka k periodiku jako celku odpovídá poli 500 v MARC21	RA	<dc:description>
<subject>	authority	údaje o věcném třídění předpokládá se přebírání z katalogizačního záznamu --- authority: vyplnit hodnotu „czenas“	R	
<topic>		libovolný výraz specifikující nebo charakterizující obsah periodika; použít kontrolovaný slovník - např. z báze autorit AUT NK ČR (věcné téma) nebo obsah pole 650 záznamu MARC21	M	<dc:subject>
<geographic>		geografické věcné třídění použít kontrolovaný slovník - např. z báze autorit AUT NK ČR (geografický termín) nebo obsah pole 651 záznamu MARC21	R	<dc:subject>
<temporal>		chronologické věcné třídění použít kontrolovaný slovník - např. z báze autorit AUT NK ČR (chronologický údaj) nebo obsah pole 648 záznamu MARC21	R	<dc:subject>
<name>		jméno použité jako věcné záhlaví použít kontrolovaný slovník - např. z báze autorit AUT NK ČR (jméno osobní) nebo obsah pole 600 záznamu MARC21	R	<dc:subject>
<classification>	authority	klasifikační údaje věcného třídění podle Mezinárodního desetinného třídění odpovídá poli 080 MARC21 ---	M	<dc:subject>

		authority: vyplnit hodnotu „udc“		
<relatedItem>	type	informace o dalších dokumentech/částech/zdrojích, které jsou ve vztahu k popisovanému dokumentu; použití pro vyjádření edice, ve které je dokument vydán, údaj o edici musí obsahovat minimálně element <title> s jejím názvem Poznámka: element <relatedItem> může obsahovat jakýkoliv jiný element MODS – jejich použití se řídí pravidly popsanými pro tyto elementy; ---------- type: hodnota „series“	RA	
<identifier>	type	údaje o identifikátorech, obsahuje unikátní identifikátory mezinárodní nebo lokální, které titul periodika má – viz přehled typů atributů níže ---------- type: budou se povinně vyplňovat následující hodnoty, pokud existují: - doi - hdl - handle - issn - převzít z katalogizačního záznamu NK ČR - isbn - převzít z katalogizačního záznamu NK ČR - ccnb – čČNB - převzít z katalogizačního záznamu NK ČR - permalink záznamu z katalogu NK ČR, např. http://aleph.nkp.cz/F/?func=direct&doc_number=002186258&local_base=NKC - uuid - jiný interní identifikátor, hodnota atributu „local“,	M	<dc:identifier>

		lze použít např. k vyjádření čárového kódu		
<location>		údaje o uložení popisovaného dokumentu, např. signatura, místo uložení apod.	MA	
<url>	note	pro uvedení lokace elektronického dokumentu --- note: pro poznámku o typu URL (na plný text, abstrakt apod.)	O	<dc:source>
<physicalLocation>	authority	údaje o instituci, kde je fyzicky uložen popisovaný dokument, např. NK ČR nutno použít kontrolovaný slovník – sigly knihoven (ABA001 atd.) odpovídá poli 040 v MARC21 pozn. u dokumentů v digitální podobě není možné vyplnit --- authority: hodnota „siglaADR“	M	<dc:source>
<shelfLocator>		signatura nebo lokační údaje o dokumentu	M	<dc:source>
<part>	type	popis částí dokumentu, bude využit jen na popis ročníku (volume) periodika --- type: hodnota bude vždy „volume“	M	
<detail>	type	upřesnění popisu části --- type: hodnota bude vždy „volume“	M	
<number>		číslo části (ročníku)	MA	<dc:description> povinné pokud lze uvést; nutno doplnit spojení „volume number“, viz <dc:description>volume number: 25</dc:description>
<caption>		text před číslem ročníku, např. „ročník“, „roč.“, „volume“ apod.	O	
<date>		datum vztahující se k části	MA	

		v případě, že se ročník vycházel během více let (přelom roku), nutno uvést oba roky, např. 1920-1921		
<recordInfo>		údaje o metadatovém záznamu – jeho vzniku, změnách apod.	M	
<recordContentSource>		kód nebo jméno instituce, která záznam vytvořila nebo změnila; nutno vytvořit kontrolovaný slovník	R	
<recordCreationDate>	encoding	datum prvního vytvoření záznamu, na úroveň minut --- encoding: záznam bude podle normy ISO 8601 na úroveň minut, hodnota atributu tedy iso8601	M	
<recordChangeDate>	encoding	datum změny záznamu --- encoding: záznam bude podle normy ISO 8601 na úroveň minut, hodnota atributu tedy iso8601	R	
<recordOrigin>		údaje o vzniku záznamu hodnoty: machine generated nebo human prepared	R	

8.4.1.2 Pole MODS a DC pro číslo periodika

Element MODS	Atributy	Popis	povinnost	Element DC
<titleInfo>	ID	název titulu periodika, kterého je číslo součástí, převzít z katalogizačního záznamu titulu periodika použít názvové autority nebo katalogizační záznam --- ID musí vyjadřovat název úrovně, tj. např. „MODS_ISSUE“	M	
<title>		názvová informace – titul periodika převzít z katalogu	M	<dc:title>
<subTitle>		podnázev periodika	RA	<dc:title>
<partNumber>		pořadové číslo vydání (čísla), např. 40;	MA	<dc:description>

		nebo u ročenek číslo určité řady/edice (část 1, řada B)		
<partName>		jméno edice nebo speciální ediční řady, např. Hygiena; lze uvést i název tematického čísla nebo zvláštního vydání; k použití u ročenek a specializovaných periodik nebo u tematických čísel nebo zvláštních vydání	R	<dc:description>
<name>	type	údaje o odpovědnosti za číslo periodika nepočítá se s vyplněním u deníků, ale např. u ročenek, zvláštních vydání čísel periodika apod., které mají vlastního autora/editora ---------- type: použít jeden z typů – personal – corporate – conference – family	MA	
<namePart>	type	údaje o křestním jméně a příjmení apod. nutno vyjádřit pro křestní jméno i příjmení ---------- type: použít jednu z hodnot: – date – doporučené pokud lze uvést – family – povinné pokud lze uvést – given – povinné pokud lze uvést – termsOfAddress – doporučené pokud lze uvést pokud nelze rozlišit křestní jméno a příjmení, nepoužije se atribut „type“ a jméno se zaznamená v podobě jaké je do jednoho	MA	<dc:creator> nutno do jednoho pole DC spojit jméno i příjmení

		elementu <namePart>		
<role>		specifikace role osoby nebo organizace uvedené v elementu <name>	MA	
<roleTerm>	type authority	popis role nutno použít kontrolovaný slovník např. z MARC21 --- type: code – kód role z kontrolovaného slovníku rolí http://www.loc.gov/marc/relators/relaterm.html) authority – údaje o kontrolovaném slovníku využitém k popisu role, k popisu výše uvedeného MARC seznamu nutno uvést authority="marcrelator"	MA	
<genre>	type	bližší údaje o typu dokumentu hodnota: issue --- type: pro upřesnění typu čísla a jednotlivých vydání povinné hodnota může být: - normal - běžné vydání - morning – ranní vydání - afternoon- odpolední vydání - evening – večerní vydání - sequence_X – pořadí vydání (sequence_1 = první vydání toho dne; sequence_2 = druhé vydání atd.) - corrected – opravené vydání - special – zvláštní vydání (např. k nějaké události) - supplement – v případě, že se příloha časopiseckého typu popisuje jako číslo	M	<dc:type>
<originInfo>		informace o původu předlohy doporučené kde lze vyplnit (např. u	RA/O	

		ročenek, kde se vydavatel měnil) **nepovinné pro deníky a běžná čísla periodik** Poznámka: Jeden nebo více výskytů elementů se předpokládá pro vydavatele, další výskyt v případě nutnosti popsat tiskaře. Pokud je nutno vyjádřit tiskaře (pole 260 podpole „f“ a „e“ a „g“ v MARC21), je nutno element <originInfo> opakovat s atributem transliteration=“printer“ a elementy <place>, <publisher>, <dateCreated>, které budou obsahovat údaje o tiskaři.		
<place>		údaje o místě spojeném s vydáním, výrobou nebo původem popisovaného dokumentu	MA	<dc:coverage>
<placeTerm>	type	konkrétní určení místa, např. Praha odpovídá hodnotě z katalogizačního záznamu, pole 260, podpole „a“ --- type – bude vždy text	MA	<dc:coverage>
<publisher>		jméno entity, která dokument vydala, vytiskla nebo jinak vyprodukovala odpovídá poli 260 podpoli „b“ katalogizačního záznamu v MARC21	MA	<dc:publisher>
<dateIssued>	qualifier	datum vydání předlohy, v případě čísla datum dne, kdy vyšlo; musí vyjádřit den, měsíc a rok, dle toho jaké údaje jsou k dispozici nutno zapsat v následujících podobách: - DD.MM.RRRR – pokud víme den, měsíc i rok vydání - MM.RRRR – pokud víme	MA	<dc:date>

		jen měsíc a rok vydání - RRRR – pokud víme pouze rok - DD.-DD.MM.RRRR – vydání pro více dní - MM.-MM.RRRR – vydání pro více měsíců --- qualifier – možnost dalšího upřesnění, hodnota „approximate“ pro data, kde nevíme přesný údaj		
<dateCreated>	qualifier	datum vytvoření předlohy bude použito pouze při popisu tiskaře, viz poznámka u elementu <originInfo> odpovídá hodnotě z katalogizačního záznamu, pole 260, podpole „g“ --- qualifier – možnost dalšího upřesnění, hodnota „approximate“ pro data, kde nevíme přesný údaj	R	
<language>		údaje o jazyce dokumentu	M	
<languageTerm>	type authority	přesné určení jazyka – kódem nutno použít kontrolovaný slovník ISO 639-2, http://www.loc.gov/standards/iso639-2/php/code_list.php --- type: použít hodnotu code authority: použít hodnotu „iso639-2b“	M	<dc:language>
<physicalDescription>		obsahuje údaje o fyzickém popisu zdroje/předlohy	M	
<extent>		údaje o rozsahu (stran, svazků nebo rozměrů); použití spíše u ročenek apod. odpovídá hodnotám v poli 300 podpolích „a“ a „c“ MARC21, pokud jsou vyplněna obě pole, bude se element <extent>	RA	<dc:format>

		opakovat; počet stránek bude vyjádřen ve fyzické strukturální mapě a bude tak vidět v aplikaci zpřístupnění i bez vyplnění tohoto pole		
<note>		poznámka o fyzickém stavu dokumentu; pro každou poznámku je nutno vytvořit nový <note> element	RA	
<abstract>		shrnutí obsahu dokumentu, zvláště pro ročenky, zvláštní vydání a tematická čísla plnit pouze v případech, že se liší od abstraktu na úrovni titulu odpovídá poli 520 MARC21	RA	<dc:description>
<note>		obecná poznámka k dokumentu odpovídá poli 500 v MARC21	RA	
<subject>	authority	údaje o věcném třídění plnit pouze pro tématická čísla, zvláštní vydání a ročenky – pouze pokud se liší od údajů v elementu <subject> na úrovni titulu --- authority: vyplnit hodnotu „czenas“	RA	
<topic>		libovolný výraz specifikující nebo charakterizující obsah čísla; použít kontrolovaný slovník - např. z báze autorit AUT NK ČR (věcné téma)	M	<dc:subject>
<geographic>		geografické věcné třídění použít kontrolovaný slovník - např. z báze autorit AUT NK ČR (geografický termín)	R	<dc:subject>
<temporal>		chronologické věcné třídění použít kontrolovaný slovník - např. z báze autorit AUT NK ČR (chronologický údaj)	R	<dc:subject>
<name>		jméno použité jako věcné záhlaví použít kontrolovaný slovník - např. z báze autorit AUT NK ČR (jméno osobní)	R	<dc:subject>

<identifier>	type	údaje o identifikátorech čísla, obsahuje unikátní identifikátory mezinárodní nebo lokální --- type: budou se povinně vyplňovat následující hodnoty, pokud existují: - doi - hdl - handle - isbn - převzít z katalogizačního záznam NK ČR (ročenky apod.) - urnnbn - pro URN:NBN, např. zápis ve tvaru urn:nbn:cz:anl-123456 pro projekt ANL+; pozor, musí odpovídat URN:NBN, podle kterého je pojmenovaný PSP balíček a jeho jednotlivé soubory - uuid - jiný interní identifikátor, hodnota atributu „local“, lze použít např. k vyjádření čárového kódu	M	<dc:identifier>
<location>		údaje o uložení popisovaného dokumentu, např. signatura, místo uložení apod. doporučené např. pro ročenky apod., kde se signatury jednotlivých čísel liší	R	
<url>	note	pro uvedení lokace elektronického dokumentu --- note: pro poznámku o typu URL (na plný text, abstrakt apod.)	O	<dc:source>
<physicalLocation>	authority	údaje o instituci, kde je fyzicky uložen popisovaný dokument, např. NK ČR nutno použít kontrolovaný slovník – sigly knihovnen (ABA001 atd.) odpovídá poli 040 v MARC21 --- authority: hodnota „siglaADR“	MA	<dc:source>
<shelfLocator>		signatura nebo lokační údaje o	MA	<dc:source>

		dokumentu		
<part>	type	popis částí dokumentu, bude využit jen na zaznamenání <caption> --- type: hodnota bude vždy „issue“	O	
<caption>		text před označením čísla, např. „č.“, „číslo“, „No.“ apod.	RA	

8.4.1.3 Pole MODS a DC pro vnitřní část periodika (článek a obraz)

Element MODS	Atributy	Popis	Povinnost	Element DC
<titleInfo>	ID	názvová informace vnitřní části --- ID musí vyjadřovat název úrovně, tj. např. „MODS_PICTURE“ pro obrázek v textu, „MODS_ARTICLE“ pro článek apod.	M	
<title>		vlastní název vnitřní části (článku, obrazu); u obrazu brát případně z popisku obrazu; pokud není titul, nutno vyplnit hodnotu „untitled“	M	<dc:title>
<subTitle>		podnázev vnitřní části (článku); za podnázev lze považovat i krátký text, který se před článkem objevuje tučným písmem (shrnutí obsahu článku)	MA	<dc:title>
<partNumber>		číslo vnitřní části např. článek na pokračování	RA	<dc:title>
<partName>		název pokračování vnitřní části (článku)	RA	<dc:title>
<name>	type	údaje o odpovědnosti za vnitřní část (článek i obraz) --- type: použít jeden z typů – personal – corporate – conference – family	MA	
<namePart>	type	údaje o křestním jméně a příjmení apod. nutno vyjádřit pro křestní jméno i příjmení	MA	<dc:creator> nutno do jednoho pole DC spojit jméno i

		---------- type: použít jednu z hodnot: – date – doporučené pokud lze uvést – family – povinné pokud lze uvést – given – povinné pokud lze uvést – termsOfAddress – doporučené pokud lze uvést pokud nelze rozlišit křestní jméno a příjmení, nepoužije se atribut „type“ a jméno se zaznamená v podobě jaké je do jednoho elementu <namePart>		příjmení
<role>		specifikace role osoby nebo organizace uvedené v elementu <name>	RA	
<roleTerm>	type authority	popis role nutno použít kontrolovaný slovník např. z MARC21 ---------- type: code – kód role z kontrolovaného slovníku rolí http://www.loc.gov/marc/relators/relaterm.html) authority – údaje o kontrolovaném slovníku využitém k popisu role, k popisu výše uvedeného MARC seznamu nutno uvést authority=“marcrelator“	MA	
<genre>	type	bližší údaje o typu vnitřní části povinné hodnota: article nebo picture ---------- type: doporučené hodnota pro article – možnost vyplnit bližší určení typu článku (možnost použít DTD periodika,	M	<dc:type>

		Article Types) - news - table of content - advertisement - abstract - introduction - review - dedication - bibliography - editorsNote - preface - main article - index (použije se pro všechny typy seznamů mimo hlavní obsah; např. seznam obrazů, tabulek apod.) - unspecified – pokud nepatří ani do jedné z výše uvedených kategorií - aj. hodnota pro picture – možnost vyplnit další určení typu obrazu - table - illustration - chart - photograph - graphic - map - advertisement - cover - unspecified – pokud nepatří ani do jedné z výše uvedených kategorií - aj.		
<language>		údaje o jazyce vnitřní části nelze plnit u obrazu	MA	
<languageTerm>	type authority	přesné určení jazyka – kódem nutno použít kontrolovaný slovník ISO 639-2, http://www.loc.gov/standards/iso639-2/php/code_list.php nelze plnit u obrazu ----------	M	<dc:language>

		type: použít hodnotu code authority: použít hodnotu „iso639-2b“		
<physicalDescription>		obsahuje údaje o fyzickém popisu zdroje/předlohy; určeno spíše pro články než pro obrazy	R	
<form>	authority	údaje o fyzické podobě vnitřní části, např. print, electronic apod. odpovídá hodnotám pozice 23 a 29 v poli 008 MARC21 --- authority: hodnota „marcform“	R	<dc:format>
<abstract>		shrnutí obsahu vnitřní části	R	<dc:description>
<note>		obecná poznámka k vnitřní části do poznámky by se měla dávat šifra autora vnitřní části, která se vyskytuje pod vnitřní částí odpovídá poli 500 v MARC21	RA	<dc:description>
<subject>		údaje o věcném třídění	R	
<topic>	authority (volitelné)	libovolný výraz specifikující nebo charakterizující obsah vnitřní části; lze (není ovšem nutno) použít kontrolovaný slovník - např. z báze autorit AUT NK ČR (věcné téma) --- při použití autoritních záznamů použít AUT NK ČR a atribut authority: vyplnit hodnotu „czenas“; při použití volných klíčových slov atribut authority nepoužívat	M	<dc:subject>
<geographic>	authority	geografické věcné třídění použít kontrolovaný slovník - např. z báze autorit AUT NK ČR (geografický termín) --- authority: vyplnit hodnotu „czenas“	R	<dc:subject>
<temporal>	authority	chronologické věcné třídění	R	<dc:subject>

		použít kontrolovaný slovník - např. z báze autorit AUT NK ČR (chronologický údaj) --- authority: vyplnit hodnotu „czenas“		
<name>	authority	jméno použité jako věcné záhlaví použít kontrolovaný slovník - např. z báze autorit AUT NK ČR (jméno osobní) --- authority: vyplnit hodnotu „czenas“	R	<dc:subject>
<classification>	authority	klasifikační údaje věcného třídění podle Mezinárodního desetinného třídění plnit pouze pro článek odpovídá poli 080 MARC21 --- authority: vyplnit hodnotu „udc“	RA	<dc:subject>
<identifier>	type	údaje o identifikátorech, obsahuje unikátní identifikátory mezinárodní nebo lokální, které vnitřní část má – viz přehled typů atributů níže --- type: budou se povinně vyplňovat následující hodnoty, pokud existují pro článek nebo obraz: - doi - hdl - handle - urnnbn - pro URN:NBN - uuid - jiný interní identifikátor, hodnota atributu „local“, lze použít např. k vyjádření čárového kódu	M	<dc:identifier> povinné
<part>		popis částí vnitřní části, bude využito na záznam rozsahu nelze u obrazu	RA	
<extent>		upřesnění popisu části – rozsah na stránkách	MA	<dc:format>
<start>		první stránka, na které vnitřní část začíná	MA	<dc:coverage>
<end>		poslední stránka, na které vnitřní	MA	<dc:coverage>

		část končí		
<recordInfo>		údaje o metadatovém záznamu vnitřní části – jeho vzniku, změnách apod.	M	
<recordContentSource>		kód nebo jméno instituce, která záznam vytvořila nebo změnila; nutno vytvořit kontrolovaný slovník	R	
<recordCreationDate>	encoding	datum prvního vytvoření záznamu vnitřní části --- encoding: záznam bude podle normy ISO 8601 na úroveň minut, hodnota atributu tedy iso8601	M	
<recordChangeDate>	encoding	datum změny záznamu vnitřní části --- encoding: záznam bude podle normy ISO 8601 na úroveň minut, hodnota atributu tedy iso8601	R	
<recordOrigin>		údaje o vzniku záznamu vnitřní části hodnoty: machine generated nebo human prepared	R	

8.4.1.4 Pole MODS a DC pro přílohu

Element MODS	Atributy	Popis	Povinnost	Element DC
<titleInfo>	ID	názvová informace přílohy použít názvové autority nebo katalogizační záznam --- ID musí vyjadřovat název úrovně, tj. „MODS_SUPPLEMENT“	M	
<title>		názvová informace – název periodika, jehož součástí příloha je převzít z katalogu	M	<dc:title>
<partNumber>		číslo přílohy, pokud nějaké má doporučené pokud lze vyplnit	MA	<dc:description>
<partName>		název přílohy	MA	<dc:title>
<name>	type	údaje o odpovědnosti za přílohu --- type: použít jeden z typů – personal – corporate – conference	MA	

		– family		
<namePart>	type	údaje o křestním jméně a příjmení apod. nutno vyjádřit pro křestní jméno i příjmení --- type: použít jednu z hodnot: – date – doporučené pokud lze uvést – family – povinné pokud lze uvést – given – povinné pokud lze uvést – termsOfAddress – doporučené pokud lze uvést pokud nelze rozlišit křestní jméno a příjmení, nepoužije se atribut „type“ a jméno se zaznamená v podobě jaké je do jednoho elementu <namePart>	MA	<dc:creator> nutno do jednoho pole DC spojit jméno i příjmení
<role>		specifikace role osoby nebo organizace uvedené v elementu <name>	MA	
<roleTerm>	type authority	popis role nutno použít kontrolovaný slovník např. z MARC21 --- type: code – kód role z kontrolovaného slovníku rolí http://www.loc.gov/marc/relators/relaterm.html) authority – údaje o kontrolovaném slovníku využitém k popisu role, k popisu výše uvedeného MARC seznamu nutno uvést authority=“marcrelator“	MA	
<typeOfResource>		popis charakteristiky typu nebo obsahu přílohy jedna z hodnot: - text – např. pro přílohu	R	<dc:type>

		typu časopis, kniha, brožura apod. - cartographic – pro mapy - notated music - sound recording-musical - pro hudební CD/DVD - sound recording-nonmusical - sound recording - still image – fotografie, plakáty apod. - moving image – pro filmová DVD - three dimensional object - software, multimedia – pro CD/DVD se SW - mixed material		
<genre>		bližší údaje o typu dokumentu hodnota: - volume_supplement (příloha k ročníku, např. obsah celého ročníku) - issue_supplement (příloha k číslu)	M	<dc:type>
<originInfo>		informace o původu přílohy plnit pokud *se liší od údajů v popisu čísla (platí i pro jednotlivé sub-elementy)* Poznámka: Jeden nebo více výskytů elementů se předpokládá pro vydavatele, další výskyt v případě nutnosti popsat tiskaře. Pokud je nutno vyjádřit tiskaře (pole 260 podpole „f“ a „e“ a „g“ v MARC21), je nutno element <originInfo> opakovat s atributem transliteration=“printer“ a elementy <place>, <publisher>, <dateCreated>, které budou obsahovat údaje o tiskaři.	MA	
<place>		údaje o místě spojeném s vydáním, výrobou nebo původem přílohy	MA	<dc:coverage>

<placeTerm>	type	konkrétní určení místa, např. Praha odpovídá hodnotě katalogizačního záznamu, pole 260, podpole „a“ --- type – bude vždy text	MA	<dc:coverage>
<publisher>		jméno entity, která přílohu vydala, vytiskla nebo jinak vyprodukovala odpovídá poli 260 podpoli „b“ katalogizačního záznamu v MARC21	MA	<dc:publisher>
<dateIssued>	qualifier	datum vydání přílohy, musí vyjádřit den, měsíc a rok, dle toho jaké údaje jsou k dispozici nutno zapsat v následujících podobách: - DD.MM.RRRR – pokud víme den, měsíc i rok vydání - MM.RRRR – pokud víme jen měsíc a rok vydání - RRRR – pokud víme pouze rok - DD.-DD.MM.RRRR – vydání pro více dní - MM.-MM.RRRR – vydání pro více měsíců možno použít hodnotu z katalogizačního záznamu, pole 260, podpole „c“ --- qualifier – možnost dalšího upřesnění, hodnota „approximate“ pro data, kde nevíme přesný údaj	M	<dc:date>
<dateCreated>	qualifier	datum vytvoření přílohy bude použito pouze při popisu tiskaře, viz poznámka u elementu <originInfo> nebo např. u popisu CD/DVD apod. odpovídá hodnotě z katalogizačního záznamu, pole 260, podpole „g“	R	

		---------- qualifier – možnost dalšího upřesnění, hodnota „approximate“ pro data, kde nevíme přesný údaj		
<frequency>		údaje o pravidelnosti vydávání odpovídá údaji MARC21 v poli 310 nebo pozici 18 v poli 008	RA	
<language>		údaje o jazyce dokumentu	M	
<languageTerm>	type authority	přesné určení jazyka – kódem nutno použít kontrolovaný slovník ISO 639-2, http://www.loc.gov/standards/iso639-2/php/code_list.php ---------- type: použít hodnotu code authority: použít hodnotu „iso639-2b“	M	<dc:language>
<physicalDescription>		obsahuje údaje o fyzickém popisu zdroje/předlohy	M	
<form>	authority	údaje o fyzické podobě dokumentu, např. print, electronic apod. povinné pro tištěné předlohy hodnota „print“, pro elektronické přílohy „electronic“ odpovídá hodnotám pozice 23 a 29 v poli 008 MARC21 ---------- authority: hodnota „marcform“	M	<dc:format>
<extent>		údaje o rozsahu (stran, svazků nebo rozměrů) odpovídá hodnotám v poli 300 podpolích „a“ a „c“ MARC21, pokud jsou vyplněna obě pole, bude se element <extent> opakovat	RA	<dc:format>
<note>		poznámka o fyzickém stavu dokumentu; pro každou poznámku je nutno vytvořit nový <note> element	RA	

<abstract>		shrnutí obsahu dokumentu odpovídá poli 520 MARC21	RA	<dc:description>
<note>		obecná poznámka k dokumentu odpovídá poli 500 v MARC21	RA	<dc:description>
<subject>	authority	údaje o věcném třídění --- authority: vyplnit hodnotu „czenas“	R	
<topic>		libovolný výraz specifikující nebo charakterizující obsah přílohy; použít kontrolovaný slovník - např. z báze autorit AUT NK ČR (věcné téma)	M	<dc:subject>
<geographic>		geografické věcné třídění použít kontrolovaný slovník - např. z báze autorit AUT NK ČR (geografický termín)	R	<dc:subject>
<temporal>		chronologické věcné třídění použít kontrolovaný slovník - např. z báze autorit AUT NK ČR (chronologický údaj)	R	<dc:subject>
<name>		jméno použité jako věcné záhlaví použít kontrolovaný slovník - např. z báze autorit AUT NK ČR (jméno osobní)	R	<dc:subject>
<classification>	authority	klasifikační údaje věcného třídění podle Mezinárodního desetinného třídění odpovídá poli 080 MARC21 --- authority: vyplnit hodnotu „udc“	M	<dc:subject>
<identifier>	type	údaje o identifikátorech, obsahuje unikátní identifikátory mezinárodní nebo lokální, které příloha má – viz přehled typů atributů níže --- type: budou se povinně vyplňovat následující hodnoty, pokud existují: - doi - hdl - handle - issn - převzít z katalogizačního záznam NK ČR - isbn - převzít	MA	<dc:identifier>

		z katalogizačního záznam NK ČR - ccnb – čČNB - převzít z katalogizačního záznam NK ČR - permalink záznamu z katalogu NK ČR, např. http://aleph.nkp.cz/F/?func=direct&doc_number=002186258&local_base=NKC - urnnbn - pro URN:NBN - uuid - jiný interní identifikátor, hodnota atributu „local“, lze použít např. k vyjádření čárového kódu		

8.5 METS část <amdSec> - Technická a administrativní metadata (MIX a PREMIS)

- technická metadata jsou určena primárně pro zachycení technických informací o formátech souborů, o výsledcích validací a kontrol
- administrativní metadata zachycují veškeré změny, procesy apod., které byly na datech i metadatech provedeny
- pro všechny typy digitálních objektů se bude využívat standard PREMIS (jeho části Object, Event a Agent); pro obrazová data pak navíc také standard MIX
- technická a administrativní metadata budou vznikat i pro prvotní sken (většinou TIFF), který se po nutných úpravách maže a dále neuchovává, metadata se ovšem uchovávají
- technická a administrativní metadata pro různé reprezentace jedné strany čísla periodika (původní TIFF, MC, ALTOXML a OCR.TXT) budou zabalena v části <amdSec> vedlejšího METS záznamu (AMD_METS.xml) ve vlastních schématech (MIX, PREMIS – části Object; Events; Agent)
    - vedlejší METS záznam (AMD_METS.xml) a je linkován z hlavního METS záznamu dokumentu
- **pro všechny reprezentace jedné strany čísla periodika bude ve vedlejším METS záznamu AMD_METS.xml existovat jedna část <amdSec>, která bude obsahovat metadata v <techMD> a <digiprovMD> podčástech pro jednotlivé soubory**
- **plnění technických metadat se předpokládá z výstupů vzniklých využitím služeb třetích stran, jako jsou JHOVE2, DROID aj.**

Část <amdSec> musí mít atribut ID a vnořený element <techMD> nebo <digiprovMD>, oba s atributem ID a vnořeným elementem <mdWrap> s atributem MDTYPE.

element	atributy	popis	Povinnost

<amdSec>	ID	element obsahující technická metadata ve standardu PREMIS nebo MIX --- ID – identifikátor konkrétní části <amdSec>, např. pro stránku 1 by hodnota mohla být „PAGE0001“	M
<techMD> nebo <digiprovMD>	ID	element rozlišující typy jednotlivých administrativních metadat --- ID pro část <techMD>: – pro části obsahující PREMIS Object hodnota „OBJ_001“ – objekt 1 (PREMIS Object pro smazaný TIFF, OBJ_002 by bylo pro MC, OBJ_003 pro ALTO XML – pro části obsahující MIX hodnota „MIX_001“ = MIX metadata pro původní TIFF, „MIX_002“ pro MC pro část <digiprovMD>: – pro části obsahující PREMIS Event hodnota „EVT_001“ apod. – pro části obsahující PREMIS Agent hodnota „AGENT_001“ apod.	M
<mdWrap>	MDTYPE	element obsahující vložené záznamy PREMIS, MIX --- MDTYPE – pro záznamy PREMIS Object, Event i Agent vždy hodnota „PREMIS“ – pro záznamy MIX hodnota „NISOIMG“	M

8.5.1 PREMIS Objects

- bude odpovídat poslední aktuální verzi v době implementace (leden 2011 - PREMIS data dictionary v. 2.1), nebo verzi předchozí
- popisovat se pomocí PREMIS Object budou soubory, tj. dle specifikace PREMIS vždy úroveň tzv. File (ne reprezentace ani bitstream)
- záznam v PREMIS Object se bude vytvářet pro následující soubory vzniklé v procesu digitalizace: 1) původní sken; 2) archivní obrazové kopie; 3) ALTO XML
- PREMIS Object se nebude vytvářet pro OCR.TXT soubory
- pro každý záznam PREMIS Object bude existovat vlastní podčást <techMD>
- záznam PREMIS Object pro jeden soubor bude obsahovat linky na Události, které jsou popsány v PREMIS Events ve stejném METS metadatovém záznamu konkrétního dokumentu (číslo, svazek) v části <digiprovMD>; vazba bude provedena přes element <premis:relatedEventIdentification>; to samé platí pro objekty, které budou nalinkovány

v případě vztahu (např. UC vznikla z MC) s popisovaným objektem přes <premis:relatedObjectIdentification>

- např. PREMIS Object záznam popisující archivní soubor JPEG 2000 je tímto způsobem nalinkován na původní sken ve formátu TIFF (resp. na jeho PREMIS Object záznam) pomocí elementu <relatedObjectIdentification>, který obsahuje ID původního objektu (např. TIFF)
- zároveň pomocí elementu <relatedEventIdentification> je záznam PREMIS Object archivního souboru JPEG 2000 nalinkován na událost, během které vznikl

– **POZOR – PREMIS Object bude vznikat a uchovávat se i pro neexistující data (původní a posléze smazaný TIFF)**

Pole záznamu PREMIS Object

Element	Popis	Povinnost	Použití pro
<objectIdentifier>	identifikátor k jednoznačnému odlišení objektu v určitém kontextu; 1-n	M	MC, XML, PS
<objectIdentifierType>	popis kontextu, ve kterém je identifikátor unikátní, např. NDK, ANL nebo název repozitáře; nutno použít kontrolovaný slovník; 1-1	M	MC, XML, PS
<objectIdentifierValue>	vlastní hodnota identifikátoru, např. img0001-master, urn.nbn.cz-123465 apod.; 1-1	M	MC, XML, PS
<objectCategory>	typ objektů, ke kterým se metadata (PREMIS Object) vztahují, např. file pro soubor, representation pro digitální reprezentaci, bitstream pro bitstream; 1-1	M	MC, XML, PS
<preservationLevel>	údaje o úrovni ochrany souboru, která se na něj vztahuje; některé soubory nejsou tak důležité jako jiné, mají menší úroveň ochrany; 0-n	M	MC, XML, PS
<preservationLevelValue>	hodnota úrovně ochrany, která je pro soubor relevantní, pro původní sken PS hodnota deleted, pro MC a XML hodnota preservation; 1-1	M	MC, XML, PS
<preservationLevelDateAssigned>	datum, kdy byla přiřazena hodnota úrovně ochrany, zápis v ISO 8601, na úroveň dne (DD-MM-RRRR) 0-1	R	MC, XML, PS
<objectCharacteristics>	technické údaje o souboru 1-n	M	MC, XML, PS

<compositionLevel>	údaj o tom, zda je nutné digitální objekt rozbalit nebo dekódovat; např. 0 (defaultně pro žádné zabalení nebo kódování); 1 pro jedno zabalení a kódovaní, podobně pak hodnota 2; 1-1	M	MC, XML, PS
<fixity>	údaje o kontrolním součtu 0-n	M	MC, XML, PS
<messageDigestAlgorithm>	použitý algoritmus kontrolního součtu, např. MD5 aj. 1-1	M	MC, XML, PS
<messageDigest>	hodnota kontrolního součtu 1-1	M	MC, XML, PS
<messageDigestOriginator>	agent (osoba, instituce, stroj, SW), který kontrolní součet vytvořil (např. JHOVE apod.) 0-1	M	MC, XML, PS
<size>	údaje o velikosti souboru v bytech 0-1	M	MC, XML, PS
<format>	údaje o formátu souboru 1-n	M	MC, XML, PS
<formatDesignation>	identifikace formátu souboru, výstup z JHOVE, PRONOM služeb apod. 0-1	M	MC, XML, PS
<formatName>	jméno formátu, např. image/tiff nebo Adobe PDF 1-1	M	MC, XML, PS
<formatVersion>	verze formátu, např. 6.0 0-1	M	MC, XML, PS
<formatRegistry>	identifikace formátu – dodatečná informace o záznamu formátů v registrech formátů (např. PRONOM aj.) 0-1	M	MC, XML, PS
<formatRegistryName>	jméno použitého registru formátů, např. UDFR, PRONOM aj. 1-1	M	MC, XML, PS
<formatRegistryKey>	unikátní identifikátor (označení) formátu v registru, např. fmt/155 z PRONOM 1-1	M	MC, XML, PS
<creatingApplication>	údaje o aplikaci, ve které byl popisovaný soubor vytvořen; nutno popsat skener, SW kde vzniklo	M	MC, XML,

	ALTO XML/TXT, SW/kodek pro vytvoření JPEG 2000 MC 0-n		PS
<creatingApplicationName>	název aplikace, např. ImageGear, Kakadu apod.; 0-1	M	MC, XML, PS
<creatingApplicationVersion>	verze aplikace, např. 15.03.000 0-1	M	MC, XML, PS
<dateCreatedByApplication>	datum a čas vytvoření, např. 2008-11-10T12:37:46; musí být ve tvaru ISO 8601 (na úroveň vteřin); 0-1	M	MC, XML, PS
<originalName>	původní jméno souboru, např. digibok_2007081301091_0011.jp2 0-1	M	MC, XML, PS
<relationship>	vyjádření vztahu popisovaného souboru k jiným souborům a událostem (eventům) 0-n	M	MC, XML
<relationshipType>	typ vztahu, doporučené hodnoty: derivation= vztah kde objekt je výsledkem změny jiného objektu; structural= vztah mezi částmi objektu; tj. např. ALTO vytvořené z TIFFU bude mít vztah derivation, podobně jako JPEG 2000 z TIFFu vytvořený; 1-1	M	MC, XML;
<relationshipSubType>	upřesnění vztahu, doporučené hodnoty: created from; has source; is source of; has sibling; has part; is part of; has root; includes; is included in; apod.; tj. např. ALTO nebo JPEG 2000 vytvořený z původního TIFFu budou mít vztah „created from“ 1-1	M	MC, XML;
<relatedObjectIdentification>	identifikace souvisejícího souboru 1-n pro MC, XML pro vyjádření vztahu k původnímu objektu (skenu)	M	MC, XML
<relatedObjectIdentifierType>	specifikace kontextu, ve kterém je identifikátor souboru jedinečný, např. URN; temporary filepath; objectID 1-1	M	MC, XML
<relatedObjectIdentifierValue>	vlastní řetězec identifikátoru, např. URN:NBN:cz-1301091_011#0001 nebo název	M	MC, XML

	souboru, cesta k souboru apod. 1-1		
<relatedEventIdentification>	identifikace s popisovaným souborem související události (eventu); seznam událostí viz PREMIS Event 0-n	M	MC, XML
<relatedEventIdentifierType>	typ události, např. interní číslovací systém událostí jako no.nb.evt; NK repository event ID, UUID apod. 1-1	M	MC, XML
<relatedEventIdentifierValue>	hodnota identifikátoru události, např. NK_EVT_005 nebo hodnota UUID aj. 1-1	M	MC, XML
<relatedEventSequence>	pořadí události, např. 003; k určení pořadí lze určit datum události 0-1	R	MC, XML
<linkingEventIdentifier>	identifikátor události týkající původního skenu PS; typy událostí mohou být např. vytvoření, smazání 0-n pro PS nutný link na události vytvoření (digitalizace) a jeho vymazání	M	PS
<linkingEventIdentifierType>	typ identifikátoru události, např. UUID, NK_eventID, vlastní číslovací systém apod. 1-1	M	PS
<linkingEventIdentifierValue>	hodnota identifikátoru, např. event_01; img0001-master-event001 apod. 1-1	M	PS

8.5.2 PREMIS Event

- bude odpovídat poslední aktuální verzi v době implementace (leden 2011 - PREMIS data dictionary v. 2.1), nebo verzi předchozí
- PREMIS Event záznamy shromažďují informace o procesech a událostech, které se týkají jednoho nebo více objektů, v našem případě souborů; primární použití je k zaznamenání událostí, které popisovaný soubor mění nebo upravují
- bude vznikat pro události, které se prováděly na obrazových datech:
    - digitalizace – vytvoření prvního skenu (např. do TIFF)
    - vytvoření ALTO XML
    - vygenerování MC
    - vygenerování UC
    - vymazání PS
- popis událostí bude zachycovat informace o jejich výsledku/výstupu

- záznamy PREMIS Event budou uloženy v METS záznamu určeném pro administrativní a technická metadata (AMD_METS.xml) v jeho části <amdSec>, podčást <digiprovMD>
- pro každou událost bude vytvořena jedna <digiprovMD> část
- každý záznam PREMIS Event je linkován na původce aktivity, tedy na PREMIS Agent záznam

Pole záznamu PREMIS Event

Element	Popis	Povinnost
<eventIdentifier>	údaje o identifikátoru události v kontextu digitalizace nebo repozitáře 1-1	M
<eventIdentifierType>	typ identifikátoru, např. no.nb.evt; NK_eventID, UUID apod. 1-1	M
<eventIdentifierValue>	hodnota identifikátoru, např. EVT_001; event_019 apod. 1-1	M
<eventType>	kategorizace události, nutno použít kontrolovaný slovník; typy událostí, které musí být zaznamenány: capture, migration, derivation, deletion 1-1	M
<eventDateTime>	datum a čas kdy byla událost provedena; nutno zapsat v ISO 8601 na úroveň vteřin 1-1	M
<eventDetail>	další údaje o události, doporučené hodnoty pro výše uvedené <eventType> následují za /: – capture/digitization – vznik prvního skenu – capture/XML_creation – capture/TXT_creation – migration/MC_creation – derivation/UC_creation – deletion/PS_deletion 0-1	M
<eventOutcomeInformation>	informace o výsledku události 0-n	R
<eventOutcome>	kategorizace výsledku události, např. slovy jako successful nebo failure, možno použít kódy – nutno používat kontrolovaný slovník nebo seznam kódů 0-1	M
<linkingAgentIdentifier>	identifikace jednoho nebo více agentů spojených s událostí 0-n	M
<linkingAgentIdentifierType>	označení typu identifikátoru, např. NK_AgentID, UUID apod. 1-1	M

<linkingAgentIdentifierValue>	hodnota identifikátoru, např. agent_softwareName_5.2; agent_novakJ apod. 1-1	M
<linkingAgentRole>	role agenta ve vztahu k události, např. software; SW component; operator; nutno používat kontrolovaný slovník 0-n	R
<linkingObjectIdentifier>	informace o objektu/souboru spojeného s událostí, link na něj 0-n	M
<linkingObjectIdentifierType>	označení typu identifikátoru, např. PhysUnitID; URN, NK_OBJ, OBJ_001 apod.; hodnoty by se měly brát z kontrolovaného slovníku 1-1	M
<linkingObjectIdentifierValue>	hodnota identifikátoru, např. URN:NBN:cz-_0011#0001 aj. 1-1	M

8.5.3 PREMIS Agent

- bude odpovídat poslední aktuální verzi v době implementace (leden 2011 - PREMIS data dictionary v. 2.1), nebo verzi předchozí
- záznam PREMIS Agent obsahuje charakteristiku tzv. agenta, který je spojen s provedenou a zaznamenanou událostí (PREMIS Event)
    - agent může být osoba, organizace nebo software
- z PREMIS Event je linkováno na agenta, který určitou akci provedl, typ ID agenta a jeho hodnota jsou uvedené v PREMIS Events (<premis:linkingAgentIdentifier>), plný popis agenta je pak v PREMIS Agent
- záznamy PREMIS Agent budou uloženy v METS záznamu určeném pro administrativní a technická metadata (AMD_METS.xml) v jeho části <amdSec>, podčást <digiprovMD>
- pro každého agenta, tj. jeden PREMIS Agent záznam, bude vytvořena jedna <digiprovMD> část
- **informace v PREMIS Event a PREMIS Object přicházející z procesu digitalizace v PSP balíčku jsou dostačující a dají nám dostatečné informace o událostech, které se odehrály v souvislosti se vznikem digitálního objektu**
    - **další upřesnění události v PREMIS Agent není nutné**

Navrhovaná pole záznamu PREMIS Agent

Element	Popis	Povinnost
<agentIdentifier>	popis identifikátoru, který jednoznačně označuje agenta v rámci jednoho kontextu (repozitář např.) 1-n	M
<agentIdentifierType>	označení typu identifikátoru, např. NK_AgentID, UUID apod.	M

	1-1	
<agentIdentifierValue>	hodnota identifikátoru, např. agent_softwareName_5.2; agent_novakJ apod. 1-1	M
<agentName>	textové upřesnění agenta, např. přesný název SW, plné jméno osoby apod. - FixImage1.3; Jan Novák; CCS docWorks 6.2.1; 0-n	R
<agentType>	obecné označení agenta – pro osoby např. osoba, pro SW např. software apod. hodnoty: organization; person; software 0-1	M
<agentNote>	použití pouze pokud je <agentType> Software a půjde o agenta souvisejícího s migrací TIFF na JPEG 2000 (creation/migration Event); bude obsahovat příkaz k výrobě JPEG 2000 souboru v aplikaci Kakadu 0-n	MA

8.5.4 Technická metadata MIX

- bude využit standard MIX, verze aktuální v době implementace projektu, nebo verze předchozí (prosinec 2010 verze 2 – viz http://www.loc.gov/standards/mix//)
- **MIX záznam vzniká pouze pro obrazové soubory**
  - **tj. bude vznikat 1) jeden záznam pro archivní kopii, 2) další záznam pro původní soubor vzniklý prvotním skenováním (nejčastěji TIFF**) a to i přesto, že tento TIFF se v průběhu výroby maže a není archivován
  - tyto dva MIX záznamy budou součástí jednoho METS záznamu AMD_METS.xml (v části <amdSec>, podčást <techMD>) pro administrativní a technická metadata, který vznikne ke každému obrazovému souboru a který je linkován z hlavního METS záznamu čísla periodika
- **MIX záznamy jednotlivých obrazových souborů se budou lišit – MIX záznam původního skenu (PS) nebude obsahovat např. element <ImageProcessing>, MIX záznam archivního souboru MC nebude naproti tomu obsahovat informace o procesu skenování, které se váží k původnímu skenu a budou v elementu <ImageCaptureMetadata> apod. – podrobnosti viz tabulka níže, sloupec „užití pro MC a PS“**
- **pro každý záznam MIX bude vytvořena vlastní část <techMD>**
- **externí služby, jako např. JHOVE a DROID, budou využívány k plnění elementů MIX**
- ve standardu MIX nebude uvedena informace o kontrolních součtech (fixity), která je obsažena v PREMIS Object a není nutno ji opakovat (viz MIX profily Nizozemí, Finska a Norska)
- element <fileSize> je pouze doporučený, údaj o velikosti souboru je součástí popisu PREMIS Object

Pole standardu MIX pro popis archivní kopie a původního skenu

Element	Popis	Povinnost	Použití pro
<BasicDigitalObjectInformation>			
<ObjectIdentifier>	údaje o identifikátoru obrazového dokumentu, který je standardem MIX popsán; 0-n	R	MC, PS
<objectIdentifierType>	např. jméno souboru, nebo jiný identifikátor; 0-1	M	MC, PS
<objectIdentifierValue>	hodnota identifikátoru, např. 20110306_001.jp2 nebo urn:nbn:123456; 0-1	M	MC, PS
<fileSize>	velikost souboru 0-1	R	MC + PS
<FormatDesignation>	údaje o formátu obrazového souboru 0-1	M	MC, PS
<formatName>	název formátu, např. lze využít MIME types[291] (Image/jp2 apod.) 0-1	M	MC, PS
<formatVersion>	verze formátu, např. 1.0 0-1	M	MC, PS
<byteOrder>	endianita, možnosti jsou little endian, middle (mix) endian a big endian 0-1	M	MC + PS
<Compression>	údaje o kompresi obrazového souboru (pokud 0-n	M	MC, PS
<compressionScheme>	informace o kompresním schématu, vyjádřeno číslem (např. 34712 je komprese JPEG 2000) nebo slovy (např. JP2 Lossless) 0-1	M	MC, PS
<BasicImageInformation>	základní technické údaje o obrazovém dokumentu 0-1	M	MC, PS
<BasicImageCharacteristics>	0-1	M	MC, PS
<imageWidth>	šířka obrazu v pixelech, např. 3987 0-1	M	MC, PS
<imageHeight>	výška obrazu v pixelech, např. 2345 0-1	M	MC, PS

[291] http://www.iana.org/assignments/media-types/index.html

<PhotometricInterpretation>	photometrická interpretace 0-1	M	MC, PS
<colorSpace>	barevný prostor, např. RGB 0-1	M	MC, PS
<ColorProfile>	údaje o barevném profilu 0-1 povinné pro dokumenty, kde je nutno uchovat přesnou reprezentaci barvy původního dokumentu a používá se ICC profil)	MA	MC + PS
<IccProfile>	ICC profil 0-1	M	MC + PS
<iccProfileName>	jméno profilu, např. sRGB, Adobe RGB aj. 0-1	M	MC + PS
<iccProfileVersion>	verze profilu, např. sRGB IEC61966-2.1 0-1	M	MC + PS
<iccProfileURI>	odkaz na profil, např. www.profily.cz/sRGB_v4_ICC_pref.icc; 0-1	R	MC + PS
<SpecialFormatCharacteristics>	speciální technické údaje o obrazovém dokumentu, použití pro formát JPEG 2000 0-1 povinný pro JPEG 2000	MA	MC
<JPEG2000>	0-1	M	MC
<CodecCompliance>	údaje o kodeku 0-1	M	MC
<codec>	název kodeku, např. Kakadu, LuraWave aj. 0-1	M	MC
<codecVersion>	verze kodeku, např. 3.1 0-1	M	MC
<codestreamProfile >	popis codestream profilu JPEG 2000, např. P0 a P1 (viz ISO/IEC 15444-4); 0-1	M	MC
<complianceClass >	specifikace největší výšky, šířky a počtu komponentů, které dekodér dokáže dekódovat, lze použít hodnoty C0, C1 a C2; 0-1	M	MC
<EncodingOptions >	obsahuje informace o kódování JPEG 2000 0-1	M	MC
<Tiles >	popis pixelové velikosti dlaždic formátu JPEG 2000 0-1	M	MC
<tileWidth>	šířka dlaždice, např. 128 0-1	M	MC

<tileHeight>	výška dlaždice, např. 128 0-1	M	MC
<qualityLayers>	číselná hodnota počtu vrstev, do kterých byl JPEG 2000 rozdělen, např. 12 0-1	M	MC
<resolutionLevels>	popis počtu nižších rozlišení, které lze z obrazu získat, např. 6 0-1	M	MC
<ImageCaptureMetadata>	popis procesu skenování, je důležité vyplnit, protože tyto údaje nelze zjistit z finálního master/archivního souboru 0-1	M	PS
<SourceInformation>	informace o předloze 0-1	R	PS
<sourceType>	Book, Newspaper aj.; nutno používat kontrolovaný slovník 0-1	M	PS
<SourceID>	identifikátor předlohy 0-n	R	PS
<sourceIDType>	typ identifikátoru, např. čČNB, URN:NBN 0-1	M	PS
<sourceIDValue>	vlastní hodnota identifikátoru 0-1 povinné	M	PS
<GeneralCaptureInformation>	základní údaje o skenování 0-1	M	PS
<dateTimeCreated>	údaj o datu a čase skenování, např. 2009-01-03T08:25:28; zapsat v ISO 8601 na úroveň vteřin 0-1	M	PS
<imageProducer>	entita provádějící skenování, např. The National Library of the Czech Republic, osoba apod. 0-1	M	PS
<captureDevice>	typ skenovacího zařízení, např. reflection print scanner; doporučené využívání hodnot z kontrolovaného slovníku 0-1	M	PS
<ScannerCapture>	údaje o skeneru 0-1	M	PS
<scannerManufacturer>	výrobce skeneru, např. 4DigitalBooks, Treventus, Zeutschel 0-1	M	PS
<ScannerModel>	údaje o konkrétním typu skeneru	M	PS

	0-1		
<scannerModelName>	jméno modelové řady skeneru, např. DL 0-1	M	PS
<scannerModelNumber>	číslo/označení modelu, např. 3000 0-1	M	PS
<scannerModelSerialNo>	výrobní číslo skeneru, např. E4R0003649 0-1	M	PS
<MaximumOpticalResolution>	údaje o maximálním optickém rozlišení skeneru 0-1	M	PS
<xOpticalResolution>	optické rozlišení na ose x, např. 300 0-1	M	PS
<yOpticalResolution>	optické rozlišení na ose y, např. 300 0-1	M	PS
<opticalResolutionUnit>	jednotka optického rozlišení, např. inch (in.) 0-1	M	PS
<scannerSensor>	popis typu snímacího senzoru skenovacího zařízení, např. matrix, linear, undefined aj. 0-1	M	PS
<ScanningSystemSoftware>	údaje o softwaru skenovacího zařízení 0-1	M	PS
<scanningSoftwareName>	název softwaru, např. Copinet 0-1	M	PS
<scanningSoftwareVersionNo>	číslo verze softwaru, např. 3.7 0-1	M	PS
<DigitalCameraCapture>	údaje o snímacím zařízení (fotoaparát) 0-1 povinné, pokud je používán fotoaparát a není používán skener	MA	PS
<digitalCameraManufacturer>	výrobce fotoaparátu, např. Canon 0-1	M	PS
<DigitalCameraModel>	popis modelu fotoaparátu 0-1	M	PS
<digitalCameraModelName>	název modelové řady, např. EOS 0-1	M	PS
<digitalCameraModelNumber>	označení modelu fotoaparátu, např. 1000D 0-1	M	PS
<digitalCameraModelSerialNo>	výrobní číslo přístroje, např. E12345 0-1	M	PS
<camerarSensor>	typ senzoru fotoaparátu, např. matrix aj. 0-1	M	PS
<CameraCaptureSettings>	údaje o nastavení fotoaparátu použitého ke snímání předloh 0-1	M	PS

<ImageData>	v rámci tohoto kontejnerového elementu budou použity následující sub-elementy: – fNumber – exposureTime – isoSpeedRatings – shutterSpeedValue – apertureValue – brightnessValue – exposureBiasValue – maxApertureValue – subjectDistance – meteringMode – lightSource – flash – focalLength – backLight – exposureIndex – sensingMethod – cfaPattern – autoFocus – PrintAspectRatio všechny hodnoty budou přebrány v případě použití fotoaparátu z údajů Exif	M	PS
<orientation>	popis orientace obrazu tak, jak je uložen vzhledem k jeho řádkům a sloupcům, např. normal*; normal, image flipper; normal, rotated 180°; unknown apod. 0-1	M	PS
<ImageAssessmentMetadata>	informace o digitálním obrazu pro jeho hodnocení a využití z hlediska dlouhodobé ochrany apod. 0-1	M	MC, PS
<SpatialMetrics>	rozměry obrázku, 2 rozměrná projekce objektů tak jak ji „vidí“ snímací zařízení 0-1	M	MC, PS
<samplingFrequencyPlane>	popis základní roviny, např. object plane (pro přímo ze předlohy digitalizované dokumenty), source object plane (pro digitalizaci mikrofilmů), camera/scanner focal plane (indikace sampl. frekvence fyzického senzoru); 0-1	R	MC + PS
<samplingFrequencyUnit>	jednotka měření sampl. frekvence, např.	M	MC, PS

	hodnoty 1= žádná pevná jednotka; 2= inch, 3=centimetr; 0-1		
<xSamplingFrequency>	údaje o počtu pixelů na jednotku samplovací frekvence pro šířku obrázku 0-1 povinné, pokud hodnota samplingFrequencyUnit je 2 nebo 3	MA	MC, PS
<numerator>	čitatel, číselné vyjádření, např. 300 0-1	M	MC, PS
<denominator>	jmenovatel, číselné vyjádření např. 1 0-1	M	MC, PS
<ySamplingFrequency>	údaje o počtu pixelů na jednotku samplovací frekvence pro výšku obrázku 0-1 povinné, pokud hodnota samplingFrequencyUnit je 2 nebo 3	MA	MC, PS
<numerator>	čitatel, číselné vyjádření, např. 300 0-1	M	MC, PS
<denominator>	jmenovatel, číselné vyjádření např. 1 0-1	M	MC, PS
<ImageColorEncoding>	doplňující údaje o barvě obrazu 0-1	M	MC, PS
<BitsPerSample>	počet bitů na kanál 0-1	M	MC, PS
<bitsPerSampleValue>	hodnota počtu bitů, např. 8, 1, 4 nebo 8,8,8 apod. 0-n POZOR – pro každou hodnotu je nutno element opakovat, tj. např. 3x element <bitsPerSampleValue> s hodnotou 8 <mix:BitsPerSample> <mix:bitsPerSampleValue>8</mix:bitsPerSampleValue> <mix:bitsPerSampleValue>8</mix:bitsPerSampleValue> <mix:bitsPerSampleValue>8</mix:bitsPerSampleValue> </mix:BitsPerSample>	M	MC, PS
<bitsPerSampleUnit>	specifikace jednotky, např. integer nebo floating point 0-1	R	MC, PS
<samplesPerPixel>	počet barevných komponentů na pixel, např. 1, 3, 4 0-1	M	MC, PS
<TargetData>	informace o kalibračních tabulkách 0-1 povinné pro obrazy, kde se dělá kontrola oproti kalibrační tabulce	MA	MC

<targetType>	typ kalibrační tabulky; 0= external (kalibrační tabulka se neobjeví na dig. obraze, je to oddělený dig. soubor); 1= internal (tabulka je naskenována spolu s předlohou a objeví se na dig. obraze); 0-n	M	MC
<targetID>	údaje o původu kalibrační tabulky 0-n	M	MC
<targetManufacturer>	výrobce/původce kalibrační tabulky, např. Eastman Kodak nebo NK ČR, oddělení kontroly kvality apod. 0-1	M	MC
<targetName>	název kalibrační tabulky, např. ColorChecker, MicrofilmScanTarget aj. 0-1	M	MC
<targetNo>	číslo nebo verze kalibrační tabulky 0-1	M	MC
<targetMedia>	údaj o tom, na jakém médiu je kalibrační tabulka, např. film, paper aj. 0-1	R	MC
<externalTarget>	údaje o externí kalibrační tabulce; např. link na http://skenservis.cz/target-00000001 nebo název a cesta ke konkrétnímu souboru 0-n povinné v případě, že byla použita externí kalibrační tabulka (targetType = 0)	MA	MC
<performanceData>	odkaz na soubor obsahující charakteristiku výkonu systému vzhledem k nastaveným hodnotám rozlišení atd.; možné hodnoty plnění – link URN nebo URL, nebo název souboru 0-n	R	MC
<ChangeHistory>	dokumentace procesů provedených na obrazovém souboru v jeho životním cyklu 0-1	M	MC
<ImageProcessing>	údaje o zpracování obrazového souboru 0-n	M	MC
<dateTimeProcessed>	2009-01-04T15:12:06; zapsat v ISO 8601 na úroveň vteřin 0-1	M	MC
<sourceData>	odkaz na původní zdrojová data, ze kterých byl vytvořen finální obrazový soubor; může to být např. URL nebo cesta do složky s původním skenem včetně názvu souboru;	M	MC

	0-1		
<processingAgency>	The National Library of the Czech Republic 0-n	R	MC

8.6 METS část <fileSec>

8.6.1 <fileSec> hlavního záznamu METS

file group

- pro obrazy i texty (ALTO XML/OCR.TXT) budou v hlavním METS záznamu použity elementy <fileGrp>, jeden element <fileGrp> bude existovat pro obrazy archivních kopií, další pro ALTO XML, další pro OCR.TXT soubory a další pro METS záznamy s technickými metadaty (AMD_METS.xml)

1. <fileGrp> pro obrazy archivních kopií, bude mít tyto atributy: ID="MC_IMGGRP" USE="Images"
   - každý soubor bude mít vlastní element <file> s následujícími atributy:
     - ID – identifikátor souboru jp2 jak je používán v METS záznamu
     - MIMETYPE – hodnota image/jp2
     - SIZE – velikost souboru jp2
     - CHECKSUMTYPE – hodnota MD5
     - CHECKSUM – hodnota kontrolního součtu
     - SEQ – pořadí souboru
     - CREATED – datum vytvoření, ISO8601 na úroveň vteřiny
   - subelementem pod <file> je element <Flocat>, který obsahuje link (ideálně v podobě nějakého identifikátoru) na obrazový soubor (xlink:href) a atribut LOCTYPE

2. <fileGrp> pro ALTO XML bude mít následující atributy: ID="ALTOGRP" USE="Layout"
   - každý ALTO XML soubor bude mít vlastní element <file> s následujícími atributy:
     - ID – identifikátor souboru ALTO XML jak je používán v METS záznamu
     - MIMETYPE – text/xml
     - SIZE – velikost souboru xml
     - CHECKSUMTYPE – hodnota MD5
     - CHECKSUM - hodnota kontrolního součtu
     - CREATED - datum vytvoření, ISO8601 na úroveň vteřiny
   - subelementem pod <file> je element <Flocat>, který obsahuje link (ideálně v podobě nějakého identifikátoru) na xml soubor obsahující ALTO (xlink:href) a atribut LOCTYPE

3. <fileGrp> pro soubory METS s technickými metadaty AMD_METS.xml bude mít následující atributy: ID="TECHMDGRP" USE="Technical Metadata"
   - každý METS xml soubor bude mít vlastní element <file> s následujícími atributy:
     - ID - identifikátor souboru AMD_METS.xml jak je používán v METS záznamu
     - MIMETYPE – text/xml

- SIZE – velikost souboru xml
- CHECKSUMTYPE – hodnota MD5
- CHECKSUM - hodnota kontrolního součtu
- SEQ – pořadí souboru
- CREATED - datum vytvoření, ISO8601 na úroveň vteřiny

– subelementem pod <file> je element <Flocat>, který obsahuje link (ideálně v podobě nějakého identifikátoru) na xml soubor AMD_METS.xml (xlink:href) a atribut LOCTYPE

4. <fileGrp> pro soubory OCR.TXT bude mít následující atributy: ID="TXTGRP" USE="Text"
   – každý OCR.TXT soubor bude mít vlastní element <file> s následujícími atributy:
     - ID - identifikátor souboru OCR.TXT jak je používán v METS záznamu
     - MIMETYPE – text/plain
     - SIZE - velikost souboru
     - CHECKSUMTYPE – hodnota MD5
     - CHECKSUM - hodnota kontrolního součtu
     - CREATED - datum vytvoření, ISO8601 na úroveň vteřiny
   – subelementem pod <file> je element <Flocat>, který obsahuje link (ideálně v podobě nějakého identifikátoru) na txt soubor (xlink:href) a atribut LOCTYPE

8.6.2 <fileSec> vedlejšího METS záznamu AMD_METS.xml

– <fileSec> ve vedlejším METS záznamu AMD_METS.xml bude obsahovat jeden element <fileGrp> s vnořenými elementy <file> pro každou reprezentaci stránky, tj. MC, ALTO XML a OCR.TXT
– atributy jednotlivých <file> elementů odpovídají atributům pro jednotlivé typy dokumentů uvedených výše pro <fileSec> hlavního METS záznamu

8.7 METS část <structMap> - Strukturální metadata a ALTO XML

8.7.1 <structMap> hlavního záznamu METS

– strukturální mapy v METS záznamu existují dvojího typu, fyzická a logická; fyzická zaznamenává hierarchické informace o dokumentu, včetně vazeb na fyzické soubory, ze kterých se skládají jednotlivé úrovně dokumentu
– 1 logická strukturální mapa v hlavním METS záznamu popisuje 1 číslo periodika a musí popisovat strukturu až na úroveň všech článků čísla
  - součástí čísla mohou být přílohy – pokud se skenují spolu s číslem, popisuje strukturální mapa METS záznamu číslo včetně přílohy (bere se jako jedno číslo)
– strukturální mapa logická i fyzická včetně linků na ALTO XML bude v hlavním METS záznamu hlavni_METS.xml
– pro každou stránku seskupuje METS logická strukturální mapa odkazy na textové bloky (nebo ilustrace), které jsou součástí té stránky
  – informace o blocích textu nebo ilustracích na stránce jsou uloženy v 1 ALTO XML souboru, který stránce odpovídá

- každý blok a každá ilustrace má unikátní identifikátor, který je použit jako odkaz v METS strukturální mapě

Vyjádření fyzické strukturální mapy

- bude mít následující atributy <structMap LABEL="Physical_Structure" TYPE="PHYSICAL">
- fyzická strukturální mapa obsahuje rodičovský , který obsahuje tyto atributy:
  - LABEL- může obsahovat titul periodika
  - TYPE – např. newspaper
  - ID – identifikátor div
  - DMDID – identifikátor části popisných metadat (možnost rozhodnout se zda na úroveň titulu nebo čísla)
- jednotlivé stránky jsou zanořeny do rodičovského elementu jako dceřiné elementy
  - pro soubory stránky bude mít tyto atributy:
    - TYPE – bude se plnit typem stránky (viz typy stránek v DTD periodika http://digit.nkp.cz/DigitizedPeriodicals/DTD/2.10/DocumentationPeriodical/Periodical.html#element_PeriodicalPage_Link031EEEA0)
    - ID – identifikátor
    - ORDERLABEL – pořadové číslo stránky, jak je na ní vytištěno
    - ORDER – pořadí stránky v čísle periodika
  - pro soubory stránky vždy obsahují link <fptr> na soubor obrazu archivní kopie, na ALTO XML, na OCR.TXT a na AMD_METS.xml pomocí elementu <par>
    - link na obrazový soubor archivní kopie má v elementu <area> následující atributy: FILEID, který obsahuje ID souboru archivní kopie
    - link na ALTO XML má v elementu <area> následující atributy: FILEID, který obsahuje ID ALTO XML souboru, dále BEGIN="P1" kde P1 je ID elementu <page> z ALTO XML souboru; a atribut BETYPE="IDREF"
    - link na OCR.TXT soubor má v elementu <area> následující atributy: FILEID, který obsahuje ID souboru OCR.TXT
    - link na AMD_METS.xml soubor má v elementu <area> následující atributy: FILEID, který obsahuje ID souboru AMD_METS.xml

Vyjádření logické strukturální mapy

- bude mít následující atributy <structMap LABEL="Logical_Structure" TYPE="LOGICAL">
- logická struktura na úroveň článků nebo např. ilustrací se popisuje pomocí do sebe zanořených elementů
- pokud stránka obsahuje jen obraz a žádný text, pak je popsána jedním elementem s atributem TYPE="PAGE" a link do souboru ALTO XML vede přímo na element <ComposedBlock>
  - lze využít jako kontejner na obrazy a další části stránky, které nejsou součástí článku
  - pro obraz je možno využít atributy a typy podřízených elementů jak je specifikováno v tabulce níže pro PICTURE, který je součástí článku
- stránky obsahující více logických oblastí jsou popsány jedním elementem, který má vnořené elementy pro každou logickou oblast, která odpovídá např. článku, ilustraci

a. pokud se jedná o jednoduchý, celistvý článek na jedné straně, tak je popsán jen jedním elementem s atributem TYPE=“article“
    - v tomto jsou dále jako další elementy zanořeny jednotlivé textové bloky (odstavce, nadpisy, obrazy apod.)
        - u každého bloku je odkaz do ALTO XML souboru na příslušný textový blok <TextBlock> – pomocí tohoto odkazu se v ALTO XML souboru nalezne jak text, tak i informace o jeho umístění na stránce (souřadnice), toto je realizováno pomocí struktury <area> v elementu <fptr>
        - u bloku tvořeného obrazem je odkaz do ALTO XML na příslušný komponovaný blok <ComposedBlock>; je realizováno pomocí struktury <area> v elementu <fptr>
        - v případě použití atributu ORDER umožňuje tento princip u článků vyjádřit i tzv. pořadí čtení jeho částí, jako jsou např. nadpis, autor, obrázek apod.

b. pokud článek není celistvý a je rozdělen na více částí, které se vyskytují na jedné nebo více stránkách, je nutné určit pořadí čtení těchto částí, opět pomocí atributu ORDER
    - pro každou část článku existuje vlastní element, podřízený hlavnímu elementu článku
    - element každé části má atribut TYPE hodnotu „article-part“ a atribut ID musí vyjadřovat, o jakou z částí se jedná, tj. např. ID=“article5-1“ odpovídá první části článku číslo pět

– do logické struktury PSP balíčku může být v případě její existence zakomponována i příloha (Supplement), která má vlastní element s atributem TYPE=“SUPPLEMENT“
    - vnořené elementy pro obraz a články i jejich použití je shodné se způsobem popisu logické struktury u elementu s atributem TYPE=“ISSUE“

Příklad

Logická mapa obsahující číslo periodika se 2 články a 1 přílohou. První článek je jen na straně jedna a má 3 součásti (titul a odstavec s normálním textem) a obrázek s popiskem i uvedeným autorem. Druhý článek začíná na straně první a jeho další odstavec je na straně druhé. První část děleného článku má titulní část a běžný text a druhá část na straně druhé obsahuje jen odstavec s běžným textem.

```
<structMap LABEL="Logical Structure" TYPE="LOGICAL">
    <div LABEL="Mladá fronta no. 5 29.06.1979" TYPE="PERIODICAL_TITLE" ID="TITLE_1" DMDID="XY">
        <div LABEL="Mladá fronta no.5 29.06.1979" TYPE="ISSUE" ID="ISSUE_1" DMDID="XY">
            <div LABEL="Boj o zrno" TYPE="ARTICLE" ID="ARTICLE_1" DMDID="XY" ORDER="0">
                <div TYPE="TITLE" ID="ARTICLE_PART_1" ORDER="1">
                    <fptr>
                        <area FILEID="ALTO_PAGE_1" BETYPE="IDREF" BEGIN="BLOCK1"/>
                    </fptr>
                </div>
                <div TYPE="NORMAL_TEXT" ID="ARTICLE_PART_2" ORDER="2">
                    <fptr>
                        <area FILEID="ALTO_PAGE_1" BETYPE="IDREF" BEGIN="BLOCK2"/>
                    </fptr>
                </div>
                <div LABEL="Obilí na poli" TYPE="PICTURE" ID="ARTICLE_PART_3" DMDID="XY" ORDER="3">
                    <div TYPE="CAPTION" ID="ARTICLE_PART_4">
                        <fptr>
                            <area FILEID="ALTO_PAGE_1" BETYPE="IDREF" BEGIN="BLOCK3"/>
                        </fptr>
                    </div>
                    <div TYPE="PICT_AUTHOR" ID="ARTICLE_PART_5">
                        <fptr>
                            <area FILEID="ALTO_PAGE_1" BETYPE="IDREF" BEGIN="BLOCK4"/>
                        </fptr>
                    </div>
                    <div TYPE="IMAGE" ID="ARTICLE_PART_6">
                        <fptr>
                            <area FILEID="ALTO_PAGE_1" BETYPE="IDREF" BEGIN="COMPOSED_BLOCK1"/>
                        </fptr>
                    </div>
                </div>
            </div>
            <div LABEL="XVI. sjezd strany" TYPE="ARTICLE" ID="ARTICLE_2" DMDID="XY" ORDER="1">
                <div TYPE="ARTICLE_PART" ID="ARTICLE_2-1" ORDER="1">
                    <div TYPE="TITLE" ID="ARTICLE_PART_1" ORDER="1">
                        <fptr>
                            <area FILEID="ALTO_PAGE_1" BETYPE="IDREF" BEGIN="BLOCK5"/>
                        </fptr>
                    </div>
                    <div TYPE="NORMAL_TEXT" ID="ARTICLE_PART_2" ORDER="2">
                        <fptr>
                            <area FILEID="ALTO_PAGE_1" BETYPE="IDREF" BEGIN="BLOCK6"/>
                        </fptr>
                    </div>
                </div>
                <div TYPE="ARTICLE_PART" ID="ARTICLE_2-2" ORDER="2">
                    <div TYPE="NORMAL_TEXT" ID="ARTICLE_PART_1" ORDER="1">
                        <fptr>
                            <area FILEID="ALTO_PAGE_2" BETYPE="IDREF" BEGIN="BLOCK1"/>
                        </fptr>
                    </div>
                </div>
            </div>
        </div>
        <div LABEL="Mladá fronta no.5 29.06.1979" TYPE="SUPPLEMENT" ID="SUPPL_1" DMDID="XY">
            ... popis článků a obrazů stejně jako u TYPE="ISSUE"
        </div>
    </div>
</structMap>
```

kde jednotlivé části obsahují a popisují…

type	Atributy	Popis	Povinnost
TITLE	LABEL TYPE ID DMDID	obsahuje údaje o titulu periodika ---------- LABEL – název titulu periodika, včetně čísla a data vydání čísla, např. Mladá fronta no. 5 29.06.1979 TYPE – hodnota „PERIODICAL_TITLE“ ID – identifikátor , např. hodnota „TITLE_1“ DMDID – obsahuje identifikátor DMD popisné části MODS titulu	M
ISSUE nebo SUPPLEMENT	LABEL TYPE ID DMDID	obsahuje údaje o čísle/příloze čísla periodika ---------- LABEL – název titulu periodika, ve stejné podobě jako u titulu, tedy např. „Mladá fronta no. 5 29.06.1979“ TYPE- hodnota ISSUE nebo SUPPLEMENT ID – identifikátor , např. hodnota „ISSUE_1“ nebo „SUPPL_1“ DMDID – obsahuje identifikátor DMD popisné části MODS čísla/přílohy	M
ARTICLE	LABEL TYPE ID DMDID ORDER	obsahující údaje o jednom článku a jeho částech ---------- LABEL – název článku TYPE – hodnota ARTICLE s pořadovým číslem, např. ARTICLE_1 ID – identifikátor elementu DMDID – identifikátor popisných metadat ORDER – pořadí článku	M
TYPE=“ARTICLE“ může obsahovat další vnořený různých typů popisující různé části článku, rozlišujeme tyto části (typy): – TITLE – SUBTITLE – AUTHOR – TRANSLATOR – NORMAL_TEXT – běžný text bez dalšího upřesnění – PICTURE – NOTE – ARTICLE_PART - u článků, které jsou rozděleny na více míst na jedné stránce nebo více stránkách - tento pro jednu součást rozděleného článku pak může obsahovat stejné části jako pro článek, tj. (TITLE, SUBTITLE, AUTHOR, TRANSLATOR, NORMAL_TEXT, PICTURE)			
TITLE	TYPE	obsahující link na textový blok s nadpisem	MA

	ID ORDER	---------- TYPE – hodnota „TITLE“ ID – identifikátor elementu, který popisuje jednu část článku (nadpis), např. hodnota „ARTICLE_PART_1“ ORDER – pořadí části článku	
<fptr> <area>	FILEID BEGIN BETYPE	FILEID – ID ALTO XML souboru, např. „ALTO_PAGE_1“ BEGIN – ID textového bloku v ALTO XML souboru BETYPE – hodnota IDREF	
SUBTITLE	TYPE ID ORDER	obsahující link na textový blok s podnadpisem ---------- TYPE – hodnota „SUBTITLE“ ID – identifikátor elementu, který popisuje jednu část článku (podnadpis), např. hodnota „ARTICLE_PART_2“ ORDER – pořadí části článku	MA
<fptr> <area>	FILEID BEGIN BETYPE	FILEID – ID ALTO XML souboru, např. „ALTO_PAGE_1“ BEGIN – ID textového bloku v ALTO XML souboru BETYPE – hodnota IDREF	
AUTHOR	TYPE ID ORDER	obsahující link na textový blok se jménem autora ---------- TYPE – hodnota „AUTHOR“ ID – identifikátor elementu, který popisuje jednu část článku (autor), např. hodnota „ARTICLE_PART_3“ ORDER – pořadí části článku	MA
<fptr> <area>	FILEID BEGIN BETYPE	FILEID – ID ALTO XML souboru, např. „ALTO_PAGE_1“ BEGIN – ID textového bloku v ALTO XML souboru BETYPE – hodnota IDREF	
TRANSLATOR	TYPE ID ORDER	obsahující link na textový blok se jménem překladatele ---------- TYPE – hodnota „TRANSLATOR“ ID – identifikátor elementu, který popisuje jednu část článku (překladatel), např. hodnota „ARTICLE_PART_3“ ORDER – pořadí části článku	MA
<fptr> <area>	FILEID BEGIN BETYPE	FILEID – ID ALTO XML souboru, např. „ALTO_PAGE_1“ BEGIN – ID textového bloku v ALTO XML souboru BETYPE – hodnota IDREF	
NORMAL_TEXT	TYPE ID ORDER	obsahující link na textový blok s běžným textem ---------- TYPE – hodnota „NORMAL_TEXT“ ID – identifikátor elementu, který popisuje jednu část článku (běžný text), např. hodnota „ARTICLE_PART_4“	M

		ORDER – pořadí části článku	
<fptr> <area>	FILEID BEGIN BETYPE	FILEID – ID ALTO XML souboru, např. „ALTO_PAGE_1“ BEGIN – ID textového bloku v ALTO XML souboru BETYPE – hodnota IDREF	
PICTURE	LABEL TYPE ID DMDID ORDER	pro obraz náležející k článku plní se, pokud se obraz vyskytuje --- LABEL – název obrazu pokud existuje TYPE - PICTURE ID – identifikátor elementu, který popisuje jednu část článku (běžný text), např. hodnota „ARTICLE_PART_5“ DMDID – link na bibliogr. popis obrazu ORDER – pořadí obrazu	MA
element s typem PICTURE může obsahovat další elementy s typy CAPTION, PICT_AUTHOR, PICT_TITLE a IMAGE; - CAPTION obsahuje text případného popisku k obrazu - PICT_AUTHOR obsahuje text se jménem případného autora obrazu - PICT_TITLE obsahuje text názvu obrazu, pokud nějaký název existuje - IMAGE – obsahuje link do souboru ALTO XML na blok popisující vlastní obraz			
CAPTION	TYPE ID	obsahující link na textový blok s popisem obrazu --- TYPE – hodnota CAPTION ID – identifikátor elementu, např. „ARTICLE_PART_6“	MA
<fptr> <area>	FILEID BEGIN BETYPE	FILEID – ID ALTO XML souboru BEGIN – ID textového bloku v ALTO XML souboru BETYPE – hodnota IDREF	
PICT_AUTHOR	TYPE ID	obsahující link na textový blok s autorem obrazu --- TYPE – hodnota PIT_AUTHOR ID – identifikátor elementu, např. „ARTICLE_PART_7“	MA
<fptr> <area>	FILEID BEGIN BETYPE	FILEID – ID ALTO XML souboru BEGIN – ID textového bloku v ALTO XML souboru BETYPE – hodnota IDREF	
PICT_TITLE	TYPE ID	obsahující link na textový blok s názvem obrazu --- TYPE – hodnota PICT_TITLE ID – identifikátor elementu, např. „ARTICLE_PART_7“	MA
<fptr> <area>	FILEID BEGIN BETYPE	FILEID – ID ALTO XML souboru BEGIN – ID textového bloku v ALTO XML souboru BETYPE – hodnota IDREF	

IMAGE	TYPE ID	obsahující link na komponovaný blok ALTO XML obsahující souřadnice vlastního obrazu --- TYPE – hodnota IMAGE ID – identifikátor elementu, např. „ARTICLE_PART_8“	MA
<fptr> <area>	FILEID BEGIN BETYPE	FILEID – ID ALTO XML souboru BEGIN – ID komponovaného bloku v ALTO XML souboru BETYPE – hodnota IDREF	
NOTE	ID	obsahující link na textový blok s poznámkami k článku --- ID – identifikátor elementu, např. „ARTICLE_PART_9“	MA
ARTICLE_PART	TYPE ID ORDER	obsahující další vnořené odkazující na jednotlivé části konkrétní části rozděleného článku; povinné pro dělený článek Pozn: pod TYPE=“ARTICLE_PART“ lze vnořit všechny typy jako pod TYPE=“ARTICLE“ --- TYPE – hodnota „ARTICLE_PART“ ID – identifikátor konkrétní části, pro první část děleného článku např. „ARTICLE_2-1“, tj. první část článku 2 ORDER – pořadí konkrétní části děleného článku	MA

Jednotlivé elementy lze kombinovat a vytvářet nové struktury.

8.7.2 <structMap> vedlejšího záznamu METS (AMD_METS.xml)

- bude obsahovat pouze fyzickou strukturální mapu (TYPE=“PHYSICAL“)
- ta bude obsahovat pouze jeden element s atributem TYPE=“PERIODICAL_PAGE“
- do budou vnořeny odkazy na jednotlivé reprezentace stránky periodika (MC, ALTO XML a OCR.TXT) pomocí elementu <fptr> s atributem FILEID

```
<structMap TYPE="PHYSICAL">
        <div TYPE="PERIODICAL_PAGE">
           <fptr FILEID="JP2_0001"/>
           <fptr FILEID="ALTOXML_0001"/>
           <fptr FILEID="OCRTXT_0001"/>
        </div>
</structMap>
```

8.8 OCR (ALTO XML a TXT OCR)

- bude použita poslední verze standardu ALTO XML aktuální v době implementace, nebo verze předchozí (prosinec 2010 verze 2 – viz http://www.loc.gov/standards/alto/)
- níže uvedená specifikace **neobsahuje všechny elementy a atributy standardu ALTO XML, obsahuje pouze ty, které jsou pro tuto konkrétní specifikaci relevantní – každý uvedený element má vyjádřenou míru relevance výrazy: povinné, doporučené a nepovinné**
- elementy a atributy, které v této specifikaci nejsou uvedeny, nejsou považovány pro účely specifikace za důležité
- ALTO XML i OCR TXT vzniknou pro všechny obrazové soubory náležející k jedné intelektuální entitě (svazku nebo číslu periodika) včetně prázdných stran, fotografií hřbetu, předsádky apod.
- ALTO XML i OCR TXT budou vznikat na úroveň stránky
- ALTO XML soubor pro zcela prázdné stránky bude obsahovat element /alto/Layout/Page/PrintSpace, ten ovšem nebude obsahovat podelementy[292]
  /alto/Layout/Page/PrintSpace/TextBlock;
  /alto/Layout/Page/PrintSpace/TextBlock/Illustration;
  /alto/Layout/Page/PrintSpace/TextBlock/GraphicalElement ani
  /alto/Layout/Page/PrintSpace/TextBlock/ComposedBlock
- struktura ALTO XML bude generovaná na úrovni rozpoznání slova generovaná OCR
- kvalita rozpoznání znaků bude akceptována do určité hranice, výstupy nebudou ručně opravovány
- struktura ALTO umožní vyhledávání textu a jeho zvýraznění na úrovni slova, pokud bude použit odpovídající prohlížeč
- obrazy reprezentující stránku, které budou použity jako UC, musí odpovídat rozměry, orientací a natočením obrazu, který byl použit pro vytvoření OCR
- OCR TXT bude vznikat z hotových ALTO XML během procesu digitalizace
- ALTO XML se bude vytvářet pouze pro novodobé dokumenty, nebo dokumenty s určitou hranicí kvality OCR
- jméno OCR souboru musí odpovídat jménu obrazového souboru, ke kterému náleží; např. pr_0007.jp2 a al_0007.xml nebo např. 123456_006_alto.xml a 123456_006_archiv.jp2
- kódování ALTO XML i TXT OCR musí být v UTF-8
- souřadnice pozic (HPOS, VPOS, WIDTH, HEIGHT) musí být vyjádřeny v pixelech
- v této specifikaci ALTO XML se počítá s OCR i pro text mimo tzv. textové „zrcadlo“, tj. mimo hlavní text, jako jsou např. čísla stránek, běžící nadpisy ani jiné části vyskytující se na okrajích stránky (top, left, top a bottom margin)
    - elementy <topMargin>, <leftMargin>, <rightMargin>, <bottomMargin> budou obsahovat elementy <TextBlock>, pro které platí stejná pravidla, jako pro element <textBlock> pro hlavní text stránky

[292] V části ALTO XML jsou kvůli větší přehlednosti uvedeny elementy spolu s cestou, která je přesně určuje v rámci schématu (XPath).

    - POZOR: údaje z OCR mimo hlavní text stránky by neměly být vyhledávatelné v aplikaci zpřístupnění, docházelo by ke zmatení uživatele a výsledků (např. při hledání titulu kapitoly by byly zobrazeny výsledky pro každou stránku, která obsahuje běžící nadpis apod.)
- pokud je na konci věty dělící znaménko, ALTO XML i OCR TXT musí obsahovat oba fragmenty slova s dělítkem a současně také kompletní slovo – je vysvětleno dále v tabulce
- ilustrace, reklamy a jiné grafické části stránky nebudou vyjádřeny v tazích /alto/Layout/Page/PrintSpace/Illustration ani Layout/Page/PrintSpace/GraphicalElement, tyto nejsou v popisu/tabulce níže vůbec uvedeny
- ilustrace, reklamy a jiné grafické části stránky budou vyjádřeny v tagu /alto/Layout/Page/PrintSpace/ComposedBlock/ s vyjádřením atributu TYPE, který bude označovat typ bloku (illustration, advertisement aj.)
    - např. ilustrace bude popsána v elementu /alto/Layout/Page/PrintSpace/ComposedBlock/GraphicalElement, kde ComposedBlock TYPE je „Illustration“
    - reklama s textem v rámečku bude popsána v elementu Layout/Page/PrintSpace/ComposedBlock/TextBlock, kde ComposedBlock TYPE je „Advertisement“
    - tabulky, grafy obdobně
- elementy /alto/Layout/Page/PrintSpace/ComposedBlock/Illustration a Layout/Page/PrintSpace/ComposedBlock/ComposedBlock také nebudou využity
- /alto/Layout/Page/PrintSpace/ComposedBlock/TextBlock a /alto/Layout/Page/PrintSpace/ComposedBlock/GraphicalElement nebudou obsahovat elementy <Shape>; tvar těchto bloků je vyjádřen v elementu <Shape> samotného elementu <ComposedBlock>; logicky pak souřadnice tvaru <TextBlock> nebo <GraphicalElement> obsaženého v /alto/Layout/Page/PrintSpace/ComposedBlock jsou většinou shodné, pokud není tvarů nebo bloků v rámci /alto/Layout/Page/PrintSpace/ComposedBlock více
- všechny vyplněné hodnoty jsou příklady plnění, plnění v konkrétní instituci je nutno specifikovat vlastními pravidly a kontrolovanými slovníky
- ALTO XML bude využíváno pro tzv. pořadí čtení, tj. článek vyskytující se na více stránkách nebo na více různých místech jedné stránky bude možné zobrazit celý a ve správném pořadí
    - k tomu je nutno znát jeho strukturu – ta bude vyjádřena v korespondujícím METS záznamu v logické strukturální mapě, která bude obsahovat odkazy na jednotlivé textové bloky článku, pomocí ID textových bloků použitých v ALTO XML

Element	Atribut	Popis	Povinnost
<Description>			
<MeasurementUnit>		měřící jednotka pro souřadnice v ALTO XML; možné hodnoty – dpi, pixel, inch1200 a mm10); inch1200 = 1/1200 inche; doporučené plnění je „mm10“	M

		nebo „pixel“; 0-1	
<sourceImageInformation>		informace o obrazovém souboru, ze kterého vzniklo ALTO XML; 0-1	M
<fileName>		jméno obrazového souboru, ze kterého bylo ALTO XML vytvářeno; ideálně i s cestou jeho uložení; např. nlaImageSeq-33386-b.tif//produkce/OCR/digibok_XY/XY_011.tiff 0-1	M
<fileIdentifier>		jedinečný identifikátor obrazového souboru; 0-n	R
<OCRProcessing>	ID	popis procesu vzniku OCR; 0-n --- ID OCR procesu, např. <OCRProcessing ID="OCRPROCES_1">; povinné	M
<preProcessingStep>		procesy před vznikem OCR, které provádí SW pro OCR (např. natočení obrazu) 0-n	M
<processingDateTime>		určení času procesu, který předcházel samotnému OCR; např. 2008-03-29T19:42:23 dle ISO 8601 na úroveň vteřin; 0-1	O
<processingAgency>		jméno nebo kód instituce, např. NK CZ, název externí firmy apod.; doporučujeme použít kontrolovaný slovník hodnot; 0-1	R
<processingStepDescription>		popis procesu (např. zarovnání, ořez apod.); 0-n	O
<processingStepSettings>		nastavení kroku popsaného v <processingStepDescription>, např. CCS OCR Processing Filter	O

		0-1	
<processingSoftware>		popis SW, který upravoval obrázek před vznikem OCR; 0-1	M
<softwareCreator>		výrobce softwaru - např. CCS Content Conversion Specialists GmbH, Germany; 0-1	M
<softwareName>		jméno softwaru - např. CCS docWORKS; 0-1	M
<softwareVersion>		verze SW, např. 6.2-1.16; 0-1	M
<ocrProcessingStep>		popis procesu vzniku OCR 1-1 – povinné pole	M
<processingDateTime>		okamžik kdy bylo OCR vytvořeno; nutno zapsat v ISO 8601 na úroveň vteřin; 0-1	M
<processingAgency>		jméno nebo kód instituce, např. NK CZ doporučujeme použít kontrolovaný slovník hodnot; 0-1	M
<processingSoftware>		popis SW, který dělal vlastní OCR; 0-1	M
<softwareCreator>		výrobce softwaru - např. ABBYY, Russia; 0-1	M
<softwareName>		jméno softwaru - např. FineReader; 0-1	M
<softwareVersion>		např. 8.0; 0-1	M
<Styles>		styly definují vlastnosti jednotlivých grafických prvků stránky. styl definovaný v elementu vrchní úrovně je použit jako výchozí pro podřízené elementy; 0-1	M
<TextStyle>	ID FONTSTYLE FONTFAMILY	definuje font textu; 0-n ----------	M

	FONTSIZE	ID pro každý text style použitý v OCR souboru – povinné FONTSTYLE – např. bold, italics apod.; doporučujeme používat kontrolovaný slovník; doporučené FONTFAMILY – např. arial, calibri apod.; doporučujeme používat kontrolovaný slovník; povinné FONTSIZE – velikost fontu, např. 10, 12 apod.; povinné	
<ParagraphStyle>	ID ALIGN	definuje formátování textových bloků; 0-n --- ID pro každý odstavec + zarovnání; např. PAR_01, PAR_02 apod. povinné ALIGN – zarovnání; povolené hodnoty: Left, Right, Center, Block aj.; povinné	M
<Layout>		layout - rozložení struktur (slov, odstavců apod.) na jedné stránce dokumentu; 1-1 povinný výskyt element není opakovací	M
<Page>	ID ACCURACY POSITION QUALITY PHYSICAL_IMG_NR HEIGHT WIDTH PC	element popisující jednu stránku dokumentu; 1-n --- ID – vygenerovaný identifikátor stránky, např. PAGE1, nebo P1 apod.; povinné ACCURACY – procentuální odhad	M

		přesnosti OCR (0-100); doporučené POSITION – pozice stránky; hodnoty k plnění: Left, Right, Foldout, Single, Cover; nepovinné QUALITY – krátký údaj o kvalitě předlohy stránky; hodnoty k plnění: OK, Missing, Missing in original, Damaged, Retained, Target, As in original; nepovinné PHYSICAL_IMG_NR - fyzické (pořadové) číslo stránky v dokumentu; vyjádřeno číslem, např. 1,2,3 apod.; povinné WIDTH – šířka stránky vyjádřená v pixelech; povinné HEIGHT – výška stránky vyjádřená v pixelech; povinné PC = Confidence level OCR souboru – hodnota mezi 0 (nejistá kvalita) a 1 (dobrá kvalita); nepovinné; pokud nevyplníte ACCURACY – tak je vyplnění doporučené	
<TopMargin>	ID HPOS VPOS WIDTH HEIGHT	horní okraj – prostor mezi vrchní hranou listu a vrchní linkou textu; 0-1 --- ID: unikátní ID pro element TopMargin, např. P1_TM0001 (page 1, topMargin0001); povinné	M

		HPOS: horizontální pozice; povinné VPOS: vertikální pozice; povinné WIDTH – šířka vrchního okraje; povinné HEIGHT – výška vrchního okraje; povinné	
<TextBlock>	stejné plnění a pravidla jako pro element <TextBlock> vnořený do elementu <PrintSpace>		MA
<LeftMargin>	ID HPOS VPOS WIDTH HEIGHT	levý okraj – prostor mezi levým okrajem stránky a textem; 0-1 ---------- ID: unikátní ID pro element LeftMargin, např. P1_LM0001 (page 1, leftMargin0001); povinné HPOS: horizontální pozice; povinné VPOS: vertikální pozice; povinné WIDTH – šířka levého okraje; povinné HEIGHT – výška levého okraje; povinné	M
<TextBlock>	stejné plnění a pravidla jako pro element <TextBlock> vnořený do elementu <PrintSpace>		MA
<RightMargin>	ID HPOS VPOS WIDTH HEIGHT	pravý okraj – prostor mezi pravým okrajem stránky a textem; 0-1 ---------- ID: unikátní ID pro element RightMargin, např. P1_RM0001 (page 1, rightMargin0001); povinné	M

		HPOS: horizontální pozice; povinné VPOS: vertikální pozice; povinné WIDTH – šířka pravého okraje; povinné HEIGHT – výška pravého okraje; povinné	
<TextBlock>	stejné plnění a pravidla jako pro element <TextBlock> vnořený do elementu <PrintSpace>		MA
<BottomMargin>	ID HPOS VPOS WIDTH HEIGHT	pravý okraj – prostor mezi spodním okrajem stránky a textem; 0-1 --- ID: unikátní ID pro element BottomMargin, např. P1_BM0001 (page 1, bottomMargin0001); povinné HPOS: horizontální pozice; povinné VPOS: vertikální pozice; povinné WIDTH – šířka spodního okraje; povinné HEIGHT – výška spodního okraje; povinné	M
<TextBlock>	stejné plnění a pravidla jako pro element <TextBlock> vnořený do elementu <PrintSpace>		MA
<PrintSpace>	ID HPOS VPOS WIDTH HEIGHT	popis tvaru pokrývajícího textové pole stránky; 0-1 --- ID: unikátní ID pro element <printSpace>, např. P1_PS0001 (page 1, printSpace0001); - povinné	M

		HPOS: horizontální pozice; povinné VPOS: vertikální pozice; povinné WIDTH – šířka textového pole; povinné HEIGHT – výška textového pole; povinné	
<TextBlock>	ID STYLEREFS HPOS VPOS WIDTH HEIGHT	popisy textových bloků na konkrétní stránce; 0-n pokud je stránka prázdná, TextBlock není potřeba uvádět; pokud je na stránce text tak ano --- ID obsahuje identifikátor textového bloku na stránce, např. "BLOCK1" nebo P1_TB0002 (stránka 1, textový blok 2); povinné STYLEREFS: reference na ID definice formátování textových bloků <ParagraphStyle>; povinné HPOS: horizontální pozice bloku; povinné VPOS: vertikální pozice bloku; povinné WIDTH – šířka textového bloku; povinné HEIGHT – výška textového bloku; povinné	MA
<Shape>		tvar textového bloku; 0-1 – pro jeden výskyt <TextBlock> jeden nebo žádný výskyt <Shape>;	RA

		plnit v případě, že je tvar textového bloku nestandardní (víceúhelník)	
<Polygon>	POINTS	popis (souřadnice) tvaru víceúhelníku; 0-1 --- POINTS – vyjádření jednotlivých bodů víceúhelníku; povinné	M
<TextLine>	ID STYLEREFS HPOS VPOS WIDTH HEIGHT	popis jedné řádky textu v rámci textového bloku; 1-n nutný alespoň jeden výskyt v rámci textového bloku --- ID obsahuje identifikátor řádky textu v textovém bloku, např. "P1_TL0002 (stránka 1, řádka 2); povinné STYLEREFS: reference na ID definice formátování textových bloků <ParagraphStyle>; nepovinné HPOS: horizontální pozice řádky; povinné VPOS: vertikální pozice řádky; povinné WIDTH – šířka řádky; povinné HEIGHT – výška řádky; povinné	M
<String>	ID CONTENT HEIGHT WIDTH HPOS VPOS CC WC	řetězec znaků – vlastní obsah OCR; znaky tvoří jednotlivá slova a více tagů <String> větu <TextLine>; 1-n v rámci <TextLine> --- ID obsahuje unikátní sekvenční číslo řetězce na stránce, např.	M

	V případě dělení slov také: SUBS_TYPE SUBS-CONTENT	"P3_ST0001" (strana 3, řetězec 1); povinné CONTENT – ukládá vlastní řetězec znaků (slovo); povinné HPOS: horizontální pozice řetězce; povinné VPOS: vertikální pozice řetězce; povinné WIDTH – šířka řetězce; povinné HEIGHT – výška řetězce; povinné CC – úroveň důvěry v přesnost OCR rozpoznání každého znaku v řetězci; jde o seznam čísel, každé z nich mezi hodnotami 0 (jistá) a 9 (nejistá) pro každý znak; např. CC=“0001“ pro CONTENT=“TEXT“; povinné WC – úroveň důvěry v přesnost OCR výstupu celého řetězce - slova (word confidence); hodnota mezi 0 (nejistá) a 1 (jistá); např. WC=“0,99“; povinné SUBS_CONTENT – obsah chybějící části řetězce v případě, že je slovo na konci řádku rozdělené i do druhého řádku; obsahuje celý řetězec - aby byl vyhledatelný i v případě, že slovo	

		se na stránce vyskytuje, ale je rozděleno; povinné SUBS_TYPE – označení typu substituce; možné hodnoty: HypPart1; HypPart2; Abbreviation; povinné - při výskytu SUBS_CONTENT *HypPart1* se vyskytuje při rozdělení slova u jeho první OCR části (u první části tagu <CONTENT> ve větě (stringu) první; *HypPart2* se vyskytuje u následujícího tagu <CONTENT> v následující větě (stringu), který obsahuje druhou část rozděleného slova/řetězce; *Abbreviation* – typ substituce používaný při rozepisování zkratek v textu na jejich plný text; při dělení slov v textu HypPart1 a HypPart2 povinné, abbreviation nepovinné	
<ALTERNATIVE>		alternativní hodnota OCR řetězce pro jednotlivá slova; 0-n lze použít v případě nejistoty rozpoznání řetězce;	O
<HYP>	CONTENT WIDTH HPOS VPOS	zápis znaku rozdělovníku slov 0-1 pro jeden výskyt <TextLine>; vždy pro poslední <String>; může se vyskytnout pouze na konci řádku (1x) --- CONTENT – obsahuje řetězec znaků, které jsou v textu použity na rozdělení slova, nejčastěji „-„; povinné WIDTH – šířka dělícího znaku;	MA

		doporučené HPOS: horizontální pozice dělícího znaku; doporučené VPOS: vertikální pozice dělícího znaku; doporučené	
<SP>	ID WIDTH HPOS VPOS	prázdný prostor mezi řádky; 0-n v rámci jednoho <TextLine>; vždy mezi řádky, tj. mezi tagy <String>; ---------- ID: unikátní ID pro prázdný prostor mezi řádky, např. P1_SP0001 (stránka 1, prázdný prostor 0001); povinné HPOS: horizontální pozice; povinné VPOS: vertikální pozice; povinné WIDTH – šířka prázdného prostoru; povinné	M
<ComposedBlock>	ID TYPE HPOS VPOS WIDTH HEIGHT STYLEREFS	blok sestávající z jiných bloků; může obsahovat PrintSpace/ComposedBlock/TextBlock, PrintSpace/ComposedBlock/Illustration, PrintSpace/ComposedBlock/GraphicalElement, /PrintSpace/ComposedBlock/ComposedBlock, tj. stejné elementy (bloky), které obsahuje samotný element /alto/Layout/Page/PrintSpace; 0-n	MA

		povinné pro vyjádření bloků textu (např. orámovaný text, reklamy), pro vyjádření ilustrací, tabulek a grafik --- ID: unikátní ID komponovaný blok, např. P6_CB0001 (stránka 6, komponovaný blok 0001); povinné TYPE – označení typu komponovaného bloku; nutné používat kontrolovaný slovník (illustration, Advertisement, apod.); povinné HPOS: horizontální pozice bloku; povinné VPOS: vertikální pozice bloku; povinné WIDTH – šířka komponovaného bloku; povinné HEIGHT – výška komponovaného bloku; povinné	
<Shape>		tvar komponovaného bloku; 0-1 – pro jeden výskyt /alto/Layout/Page/PrintSpace/ComposedBlock jeden nebo žádný výskyt /alto/Layout/Page/PrintSpace/ComposedBlock/Shape; doporučeno – v případě, že je tvar komponovaného bloku nestandardní (víceúhelník)	RA
<Polygon>	POINTS	popis tvaru víceúhelníku; 0-1 --- POINTS – vyjádření jednotlivých	M

		bodů víceúhelníku povinné	
<TextBlock>	ID STYLEREFS HPOS VPOS WIDTH HEIGHT	v případě, že komponovaný blok (např. orámovaný tvar) obsahuje text; platí stejná pravidla jako pro normální element /alto/Layout/Page/PrintSpace/TextBlock; 0-n (pro jeden výskyt <ComposedBlock> 0 nebo více elementů /alto/Layout/Page/PrintSpace/ComposedBlock/TextBlock>; plnit pokud je v komponovaném bloku text ---------- ID obsahuje identifikátor textového bloku v komponovaném bloku, např. P1_CB0002_SUB (stránka 1, textový blok 2, SUB značí komponovaný blok); povinné STYLEREFS: reference na ID definice formátování textových bloků /alto/Styles/ParagraphStyle; povinné HPOS: horizontální pozice bloku; povinné VPOS: vertikální pozice bloku; povinné WIDTH – šířka textového bloku; povinné HEIGHT – výška textového bloku; povinné	MA
<TextLine>	/alto/Layout/Page/PrintSpace/ComposedBlock/TextBlock/TextLine a ostatní elementy v rámci		

	/alto/Layout/Page/PrintSpace/ComposedBlock/TextBlock mají stejná pravidla a výskyty jako jako ve vrchním elementu /alto/Layout/Page/PrintSpace/TextBlock		
<GraphicalElement>	ID HPOS VPOS WIDTH HEIGHT	popis grafického tvaru; v případě využití v rámci /alto/Layout/Page/PrintSpace/ComposedBlock označuje rozměry tvaru v rámci něhož je tabulka, ilustrace, reklama apod.; 0-1 - pro jeden výskyt /alto/Layout/Page/PrintSpace/ComposedBlock 0 nebo max. 1 výskyt <GraphicalElement>; plní se pokud je na stránce a tedy v komponovaném bloku ilustrace, tabulka apod.; ---------- ID – identifikátor grafického tvaru; povinné HEIGHT – výška grafického tvaru; povinné WIDTH – šířka grafického tvaru; povinné HPOS – horizontální pozice grafického tvaru; povinné VPOS – vertikální pozice grafického tvaru; povinné	MA

11.2 Aplikační metadatový profil pro digitalizaci monografií

Definice metadatového profilu pro digitalizaci monografií v projektu NDK

VERZE 0.3 - 12.2.2012

1. Východiska

- uživatelské kopie = UC
- archivní kopie = MC
- původní sken = PS – obrazový soubor vzniklý při digitalizaci, který se po zpracování (ořez, narovnání apod.); maže se ještě v procesu digitalizace a dále se neukládá
- u všech metadatových standardů budou použity verze aktuální v době implementace projektu NDK, nebo verze předchozí v případě, že nová verze je nová min. 3 měsíce
- základní intelektuální entita ve workflow digitalizace a následně i v LTP systému = svazek monografie
- PSP balíček – *producer submission package*
    - balíček dat a metadat, který přichází od producenta dat (z workflow digitalizace)
    - PSP balíček bude obsahovat kompletní intelektuální entitu tj. svazek monografie
    - z workflow digitalizace lze poslat více PSP balíčků v balíku např. [.tar] apod.
    - pokud má vícesvazkové dílo v katalogu knihovny bibliografický záznam pro každý svazek, vznikne pro každý svazek PSP balíček a každý svazek bude brán jako jedna intelektuální entita; to samé platí i pro případ, že vícesvazkové dílo má pouze jeden záznam
- v běžném workflow digitalizace monografií se nebude provádět členění na vnitřní části (kapitoly apod.), pouze u některých zvláště důležitých monografií; pro tento případ musí existovat možnost vyjádřit popis částí (např. kapitoly, přílohy apod.) v metadatech
- základní bibliografická metadata budou stahována do workflow digitalizace z knihovních katalogů
- jako výchozí SW pro vytváření souborů JPEG 2000 se bude používat Kakadu
- úpravy obrazu, které vedou ke změně rozměrů obrazu, rozlišení apod., se musí dělat před tím, než se udělá OCR, tj. budou se dělat na TIFF souborech
- veškerá metadata musí pro zápis používat kódování UTF-8

2. Výstupy digitalizace

1. archivní kopie (1 MC pro každou stránku)
2. uživatelské kopie (1 UC pro každou vzniklou MC, tedy stránku)
3. OCR – ALTO XML soubor pro každou stránku
4. OCR TXT soubor – pro možnost stáhnout si jen text dokumentu (tam kde kvalita OCR je odpovídající), vyhledávání/indexace.
5. metadata pro MC
    - bibliografická metadata – MODS a DC
    - strukturální metadata – METS
    - technická metadata – MIX, PREMIS
    - administrativní metadata – PREMIS, METS
6. kontrolní metadatové soubory (s kontrolními součty a údaji o vzniku dat apod.)

Pozn.
Záznam METS nebude obsahovat žádná metadata pro uživatelské kopie. METS neobsahuje popisná, ani technická metadata pro UC. Obrazové soubory UC nejsou ani součástí strukturální mapy <structMap> ani <fileSec>. Součástí PSP balíčku budou jen obrazy UC ve složce [userCopy]. Důvodem je to, že metadata pro UC budou vytvářena na vstupu do Krameria4 ve standardu FOXML (Fedora Object XML). Budou se vyrábět z METS záznamu pro MC, jehož specifikace je níže.

3. Granularita metadatového záznamu

Monografie

- základní intelektuální entitou pro monografie je 1 svazek
- pokud má monografie pouze jeden svazek, vznikne jeden metadatový popis (METS záznam)
- pokud má monografie svazky dva, např. dvousvazkový slovník, jedná se o dvě intelektuální entity (svazek první a svazek druhý) a vzniknou tedy dva metadatové záznamy; ke každému svazku jeden METS záznam a tedy dva PSP balíčky
- v knihovních katalozích jsou někdy vícesvazkové monografie katalogizovány jako jeden soubor, tj. mají jeden souborný záznam v katalogu; někdy jsou jednotlivé díly vedeny jako jednotlivé záznamy v katalogu; v obou případech musí vzniknout metadatový popis ke každému svazku jako základní intelektuální entitě a také PSP balíček pro každý svazek
    - jednotlivé METS záznamy svazků, které k sobě logicky patří (vícedílné dílo) budou každý obsahovat v základním bibliografickém popisu shodné údaje vycházející ze souborného katalogizačního záznamu, včetně identifikátoru (např. čČNB); tyto údaje poté umožní logicky oba svazky spojit a prezentovat uživateli

4. Identifikátory

Do workflow digitalizace budou přicházet bibliografická metadata, která již mohou obsahovat následující identifikátory vrchních úrovní intelektuálních entit (úroveň titulu):

- ISBN – pro titul monografie (jednosvazkové); nebo pro soubor monografií, které mají pouze jeden souborný záznam, ISBN není přiděleno vždy;
- ISSN – pro titul periodika, ISSN není přiděleno vždy (chybí např. u starých titulů z 19. století);
- čČNB – identifikátor entity tak jak odpovídá katalogizačnímu záznamu, tj. každá entita se záznamem v katalogu NK/MZK má tento identifikátor.

Pokud není dostupný ani jeden z výše uvedených, lze použít čárový kód dokumentu, systémové číslo, signaturu, nebo systémové číslo kombinované s polem 001 MARC záznamu apod. Jednou z možností je také využití jednoznačného čísla UUID.

Svazkům monografie bude během digitalizace[293] přidělován identifikátor URN:NBN. Stejný identifikátor může být přidělen také nižším logickým úrovním (entitám) – tedy vnitřní část (např. článek ve sborníku), samostatná příloha apod. Syntax URN:NBN musí odpovídat specifikaci identifikátoru URN:NBN pro resolver NK ČR (např. urn:nbn:cz:ndk-123456 pro výstupy z projektu NDK).

[293] Buď přímo v SW pro workflow digitalizace, nebo za pomoci aplikace Resolver URN:NBN.

5. Struktura PSP balíčku

Níže je podoba struktury balení dat a metadat v jednom PSP balíčku na výstupu z workflow digitalizace. PSP balíček = 1 složka pro svazek monografie. V případě, že má monografie 2 svazky/díly, tak 1 svazek = 1 PSP.

složka	obsahuje >	obsahuje >	obsahuje>	
svazek monografie /číslo periodika				
	info.xml	údaje o vzniku balíku		
	složka [masterCopy]	obrazy JPEG2000 lossless		
	složka [userCopy]	obrazy JPEG2000 lossy		
	složka [ALTO]	soubory ALTO XML		
	složka [TXT]	soubory OCR.TXT		
	složka [amdSec]	AMD_METS. xml soubor pro každou stránku obsahuje>	amdSec	**techMD** = PREMISobject pro MC, původní TIFF, ALTO XML) + MIX pro MC, původní TIFF)
				digiprovMD = PREMISevent + PREMISagent
			fileSec	odkazuje na MC, ALTO XML, OCR TXT soubor popisované 1 stránky
			StructMap	pouze fyzická - pro soubory popisované stránky (MC a ALTO XML, OCR TXT)
	Hlavni_METS.xml	dmdSec	MODS a DC pro jednotlivé úrovně dokumentu	
		fileSec	obsahuje linky na MC, ALTO XML, OCR TXT a technická metadata ve složce [amdSec]	
		structMap (včetně ALTO odkazů)	logická a fyzická pro MC, ALTO XML areas, OCR TXT a AMD_METS.xml	
	MD5	kontrolní součty všech souborů v PSP balíku		

Jedná se o variantu, kdy technická a administrativní metadata nejsou obsažena v hlavním METS záznamu, ale pro každou stránku v jiném dalším METS záznamu (AMD_METS.xml). Důvodem je to, že pokud by bylo vše v hlavním METSu, byl by neúměrně dlouhý. Takto je z hlavního záznamu vedlejší METS nalinkován.

Hlavní složka PSP balíčku obsahuje následující složky a soubory:

soubor info.xml

Velmi krátce tu budou zaznamenány údaje o vzniku celého PSP balíčku – kdo a kdy ho vytvořil, jakou měl velikost, odkud kam byl nakopírován apod. Obsahovat by také měl informaci o stavu zpracování balíčku. Zaznamenány by také měly být údaje o obsahu PSP balíčku – počet a názvy souborů apod. Soubor info.xml by také mohl být vedle hlavního PSP balíčku. Údaje a struktura info.xml souboru:

1. vznik balíčku – datum dle ISO8601 na úroveň vteřin
2. ID balíčku – použít identifikátor čísla periodika (URN:NBN) – viz názvová konvence v kapitole 6
3. ID titulu - čČNB, ISBN nebo ISSN
4. údaje o větším celku, do kterého balíček patří - např. digitalizace pro ANL
5. název instituce, která je zadavatelem digitalizace
6. tvůrce balíčku – kód instituce (firmy), která balíček vytvořila
7. velikost balíčku – v kB
8. z jakého serveru bylo nahráno – URL
9. obsah balíčku
    - názvy souborů včetně directory path a koncovky (MIME type)
    - počet souborů v balíčku celkem
10. stav zpracování – možné hodnoty
    - hotovo
    - opraveno
        - added OCR
        - added titles
        - added logical parts (issues, years)
        - updated xml (Mods, DC, identifikátory),
11. poznámka – např. o tom, že balíček neobsahuje OCR apod.

Příklad balíčku, který obsahuje 2 soubory, jeden v kořenu složky a druhý ve složce:

1. CREATED=2009-11-10T12:37:46
2. PACKAGEID=ANL_123456
3. TITLEID=ISBN12341236
4. COLLECTION=ANL
5. INSTITUTION=NKP
6. CREATOR=NazevFirmy
7. PACKAGESIZE=36000155kb
8. SOURCELOCATION= server123.firma.cz/baliky_hotovo/01/2011/12/000025456
9. ITEMLIST=scan01.jp2
   ITEMLIST=slozka/hotovo/27.9.2011/scan02.jp2
   ITEMTOTAL=2
10. STATUS=hotovo
11. NOTE=noOCR

složka [masterCopy]

Složka s archivními kopiemi, obsahuje soubory JPEG 2000 v neztrátové kompresi, 1 soubor = 1 stránka, tj. obsahuje všechny naskenované stránky monografie.

složka [userCopy]

Složka s uživatelskými kopiemi, pro každou naskenovanou stránku monografie obsahuje jeden JPEG 2000 soubor se ztrátovou kompresí.

složka [ALTO]
Obsahuje ke každé stránce 1 ALTO XML soubor, tj. tolik ALTO XML souborů kolik je stránek svazku monografie.

složka [TXT]
Obsahuje ke každé stránce 1 OCR soubor jako čistý text. Tj. tolik OCR.TXT souborů, kolik je stránek svazku monografie.

složka [amdSec]
Složka s technickými metadaty – **obsahuje pro každou naskenovanou stránku monografie 1 METS soubor/záznam (AMD_METS.xml)**. Záměrně nejsou tato metadata v hlavním METS záznamu (hlavni_METS.xml), musí z něj být ovšem nalinkována (z části fileSec). Každý METS záznam AMD_METS.xml obsahuje následující části METS standardu:

- amdSec – administrativní metadata – obsahuje část:
  - technických metadat (techMD), která ve standardu PREMIS Object popisuje vlastnosti archivních kopií, ALTO XML, původního TIFF souboru, ze kterého vznikly archivní kopie. Dále je přítomen záznam technických metadat v MIX standardu pro archivní kopie a pro původní TIFF.
  - metadat o provenienci digitálních objektů (digiProvMD) – v této části je využit standard PREMIS Event a PREMIS Agent.
  - fileSec – sekce s odkazy na soubory. V případě tohoto METS záznamu pro jednu stránku, který vzniká primárně k zachycení technických a administrativních metadat, bude odkazovat na soubory, které jsou s tou konkrétní stránkou spojeny, tj. archivní kopie, ALTO XML a OCR TXT. Jde o povinnou sekci METS záznamu, pro další mapování do LTP systému nebude potřeba.
  - structMap – pouze fyzická strukturální mapa, povinná část METS záznamu. Bude ukazovat strukturu souborů k dané stránce, tj. opět archivní kopie, ALTO XML a OCR TXT. Pro další mapování do LTP systému nebude potřeba.

soubor Hlavni_METS.xml
Další částí PSP balíčku je hlavní METS dokument. Hlavní METS záznam obsahuje:

- dmdSec – bibliografická metadata k svazku monografie včetně popisu nadřazených entit (např. titul) nebo naopak částí (např. kapitola). Základ bude z katalogu, případný další popis logických částí vznikne v procesu digitalizace. Hlavním standardem metadat je MODS, nutná pro LTP je i přítomnost zkráceného záznam v Dublin Core.
- fileSec – hlavní část s linky na všechny digitální objekty (archivní kopie, ALTO XML a OCR TXT), které se váží k jednomu svazku monografie. Obsahuje také linky na administrativní metadata AMD_METS.xml do složky [amdSec].
- structMap – strukturální mapa pro celý dokument, tj. pro jeden svazek monografie. Obsahuje:
  - logickou část – vyjadřuje logickou strukturu svazku s odkazy na ALTO XML;
  - fyzickou část obsahující informace o všech reprezentacích konkrétní stránky (archivní kopie, ALTO XML, OCR TXT a AMD_METS.xml);
  - mapování na ALTO XML areas.

soubor MD5

Poslední částí PSP balíčku je soubor s kontrolními součty pro všechny soubory balíčku (kromě info.xml a .md5 souboru samotného). Soubor .md5 je jeden pro 1 celý balíček PSP (balíček se svazkem monografie). Tento soubor .md5 obsahuje kontrolní součet pro každý soubor obsažený v PSP balíčku. Z tohoto důvodu nejsou samostatné kontrolní součty součástí podsložek balíčku. Kontrolní součty jsou také samozřejmě v technických metadatech.

6. Názvová konvence složek a souborů

pojmenování PSP balíčku

- každý PSP balíček přicházející z digitalizace by měl obsahovat pouze jedinou intelektuální entitu (svazek monografie). **Pak musí název balíčku vycházet z identifikátoru této entity, např. URN:NBN, číslo čárového kódu použitého na fyzické jednotce apod.**
- **každý svazek monografie musí mít svůj jednoznačný identifikátor, tím pádem pak každý PSP balíček a každý soubor v něm má vlastní jednoznačný identifikátor**

pojmenování složek

- viz návrh struktur PSP balíčku (kapitola 5)

pojmenování souborů

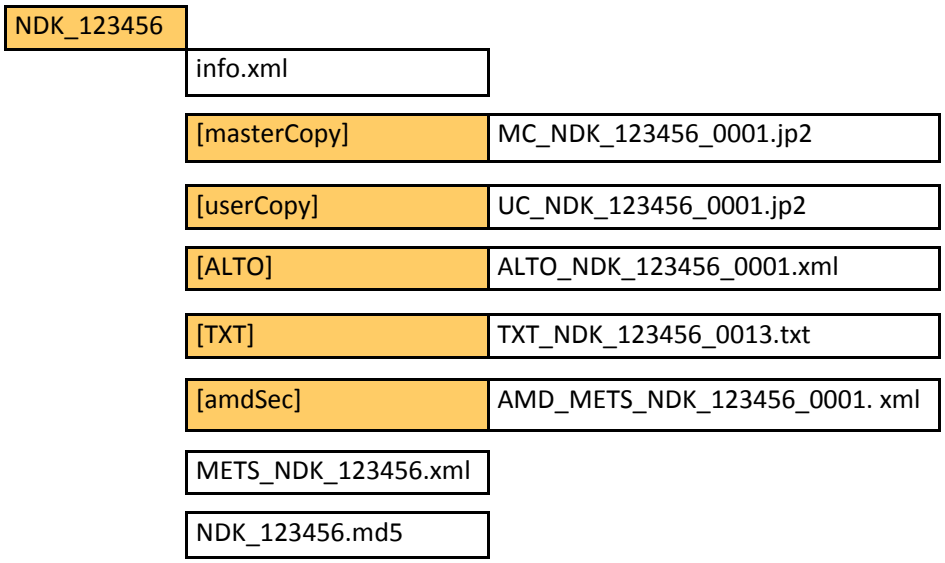
- názvy jakýchkoliv souborů náležejících k jedné základní entitě (svazek) musí být založeny na jednom typu identifikátoru
- pro svazek monografie by takovým identifikátorem mohlo být URN:NBN, čČNB, ISBN nebo ISSN titulu
- podobně využitelným identifikátorem by mohlo být generované číslo UUID, které by se generovalo pro každý soubor; tím by se ovšem ztratila (i vizuální) vazba na vrchní úroveň titulu i vazba na související soubory (stránka v jp2 a k ní náležející soubor ALTO XML apod.)

S využitím URN:NBN to může vypadat následovně (použit příklad pojmenování pro projekt NDK – digitalizace monografií):

typ souboru	název souboru	vysvětlení
PSP balíček (číslo, svazek)	NDK_123456	název celé složky PSP balíčku, u základních intelektuálních entit bude v názvu využito vždy URN:NBN
archivní kopie	MC_NDK_123456_0013.jp2	archivní JPEG 2000 stránky 13 svazku monografie s urn:nbn:cz:ndk-123456
uživatelská kopie	UC_NDK_123456_0013.jp2	uživatelská kopie ve formátu JPEG 2000 stránky 13 svazku monografie s urn:nbn:cz:ndk-123456
ALTO XML	ALTO_NDK_123456_0013.xml	ALTO soubor náležející ke 13té stránce ze svazku monografie s urn:nbn:cz:ndk-123456

OCR TXT	TXT_NDK_123456_0013.txt	TXT soubor s OCR náležející ke 13té stránce ze svazku monografie s urn:nbn:cz:ndk-123456
info.xml	INFO_NDK_123456.xml	informační XML k celému PSP balíčku svazku monografie
MD5	NDK_123456.md5	soubor s kontrolními součty k celému PSP balíčku svazku monografie
Hlavni_METS.xml	METS_NDK_123456.xml	hlavní METS záznam k celému svazku monografie s urn:nbn:cz:ndk-123456
AMD_METS.xml	AMD_METS_NDK_123456_0013.xml	METS záznam s technickými metadaty pro stránku 13 ze svazku monografie s urn:nbn:cz:ndk-123456

Složka jednoho balíčku PSP, který obsahuje jen jeden obrazový soubor k první stránce svazku monografie, pak může vypadat následovně (příklad balíčku z digitalizace NDK):



7. Transportní balík pro jeden nebo více PSP balíčků

Pokud bude jeden PSP balík obsahující 1 základní intelektuální entitu (svazek monografie) přemísťován např. jako tar soubor, měl by název souboru tar odpovídat názvu PSP balíčku (tedy vycházet z použitého identifikátoru pro entitu svazku).

Výstupem workflow digitalizace ale může také být balík (např. tar), který obsahuje více PSP balíčků. Toto sdružování bude omezeno jen kapacitou HW. Takovýto sdružený balík by měl být pojmenován na základě již užívaného identifikátoru.

- v případě, že balík obsahuje svazky jednoho vícesvazkového díla, měl by název balíku vycházet ze společného čČNB nebo ISBN, pokud existují
- typ identifikátoru musí být vyjádřen v názvu souboru – např. ISBN_1234567890.tar nebo CCNB_12345678910.tar apod.

- lze počítat s tím, že bude docházet k tomu, že sdružený balík nebude obsahovat např. všechny svazky titulu monografie – tato skutečnost musí být patrná z názvu balíku (např. ISBN_1234567890_YYYY; kde YYYY může být pořadové číslo, datum, doba vzniku jednoho z více balíků obsahujících svazky určitého titulu/souboru s identifikátorem ISBN 1234567890)

Transportní balík by měl obsahovat následující části:
- balíčky PSP (svazků);
- informační soubor, který odpovídá specifikaci info.xml;
- kontrolní součty všech PSP balíčků;
- seznam balíčků v transportním balíku.

8. Metadata

- veškerá metadata budou „zabalena“ pomocí kontejnerového standardu METS
- standard METS bude v aktuální verzi v době implementace nebo verzi předchozí (prosinec 2010 verze 1.9- http://www.loc.gov/standards/mets/mets-schemadocs.html)
- veškerá metadata ve všech standardech musí být zapsána pomocí XML za použití kódování UTF-8
- **vložení metadatových standardů do kontejneru METS bude vždy formou <mdWrap>, tj. ne odkazováním z METS záznamu ven**

8.1 Vysvětlivky k tabulkám

Obsah pole „Použití pro“
použití jednotlivých elementů pro popis MC, PS (původní sken), XML (ALTO)

Pole „Popis“ obsahuje:
- vysvětlení a příklad
- doporučené plnění tam, kde je to možné uvést
- vysvětlení a doporučené hodnoty atributů
- opakovatelnost výskytu elementu (např. pro standard PREMIS – dle XSD)
    - 0-1 element je nepovinný, neopakovatelný
    - 0-n element je nepovinný, opakovatelný
    - 1-n element je povinný a opakovatelný
    - 1 element je povinný a neopakovatelný

Význam pole „povinnost“
Pole „povinnost“ uvádí, zda je plnění jednotlivých elementů povinné, doporučené nebo volitelné. Může nabývat následujících hodnot:
- M mandatory (povinně plnit – element je součástí každého záznamu)
- MA mandatory if available (povinně plnit pokud je to možné, pokud lze apod.)
- R recommended (plnění hodnot elementu je doporučeno, není ovšem povinné)
- RA recommended if available (doporučeno pokud lze plnit)
- O optional (plnění hodnot elementu je zcela dle konkrétních potřeb)

- v případě tabulky obsahující dvě schémata (např. MODS a Dublin Core) platí povinnost pro elementy obou schémat stejně
- pokud je rodičovský element např. doporučený a dceřiný element povinný, znamená to, že dceřiný element je povinný pouze tehdy, pokud je použit element rodičovský
- oranžová barva v tabulkách označuje elementy, které mají povinný výskyt

8.2 Kořenový element hlavního METS záznamu

element	atributy	popis	Povinnost
<mets>	LABEL TYPE	kořenový element METS záznamu ---------- LABEL – název titulu monografie, včetně roku vydání, např. Honzíkova cesta, 1979 TYPE – hodnota vždy „Monograph“	M

Kořenový element hlavního METS záznamu k jednomu svazku monografie musí obsahovat linky na specifikace jednotlivých použitých metadatových schémat (METS, MODS, Dublin Core).

8.3 METS hlavička <metsHdr>

Dokumentuje vznik a úpravy METS záznamu.

element	atributy	popis	povinnost
<metsHdr>	LASTMODDATE CREATEDATE	hlavička METS záznamu ---------- LASTMODDATE – datum poslední úpravy záznamu, musí být ve tvaru ISO 8601 (na úroveň vteřin) CREATEDATE – datum vytvoření záznamu, musí být ve tvaru ISO 8601 (na úroveň vteřin)	M
<agent>	ROLE TYPE	údaje o tvůrci záznamu METS ---------- ROLE – hodnota „CREATOR“ TYPE – hodnota „ORGANIZATION“	M
<name>		jméno jednotlivce nebo organizace; ---------- tvůrce záznamu, buď dodavatel (firma XY) nebo v případě tvorby záznamu v knihovně bude využita sigla knihoven, tj. pro NK ČR hodnota „ABA001“	M

8.4 METS část <dmdSec> - Bibliografická metadata monografií (MODS a Dublin Core)

- na samotný bibliografický popis bude použit standard MODS, aktuální verze v době implementace, nebo verze předchozí (prosinec 2010 verze 3.4 viz (http://www.loc.gov/standards/mods/) a nekvalifikovaný Dublin Core (DC) (http://dublincore.org/documents/dcmi-terms/)
- DC je primárně určeno na poskytnutí dat přes OAI-PMH, bude odpovídat OAI XSD (viz http://www.openarchives.org/OAI/2.0/oai_dc.xsd)
- DC bude uloženo v METS apod. stejným způsobem jako standard MODS – viz možnosti struktury PSP balíčku výše
- pro vytvoření DC z MODS standardu může být použito oficiální mapování Kongresové knihovny – viz http://www.loc.gov/standards/mods/mods-conversions.html
- DC a MODS bude vložen v METS části dmdSec – viz možnosti struktur PSP balíčku v kapitole 5
- základním zdrojem pro popisná metadata je katalog NKČR a MZK
- u digitalizovaných dokumentů je bibliografický popis vytvářen primárně z pohledu popisu fyzické předlohy, nejde o popis elektronického dokumentu

Základní intelektuální entitou pro popis je svazek monografie, tj. v jednom METS záznamu, který bude obsahovat metadata a strukturu jednoho svazku, budou MODS záznamy k tomuto svazku.
Metadata budou popisovat entity[294]:

- ***svazek (Volume)***
- ***vnitřní část (InternalPart)*** – typy „textový oddíl“ (Chapter) a „obraz“ (Picture)
- ***příloha (Supplement)***

Pozn:
Pro popisná metadata monografií profilu NDK se nepočítá s úrovní souborného titulu, a to ani pro vícedílné publikace, které mají pouze jeden katalogizační (souborný) záznam. U jednosvazkových monografií titul splývá s popisem svazku (MODS záznam popisující svazek je záznam titulu z katalogu NK ČR/MZK). Pokud přeci jen existuje souhrnný název pro více svazků (např. sebrané spisy), je řešeno plněním sournného názvu do údajů o edici.

- záznamy monografie mohou být v katalozích následující:
    1. monografie má jen jeden svazek – existuje jeden záznam v katalogu a jedno čČNB
    2. monografie má více svazků – pak existuje buď:
        - jeden záznam pro soubor, pokud jednotlivé svazky/díly nejsou od sebe příliš odlišné (např. slovník A-K, L-Z), k jednomu záznamu existuje jedno čČNB; nebo
        - v případě, že jednotlivé díly/svazky souboru jsou odlišné (např. Vlastivěda česká – díl Flora, díl Fauna, atd.), tak má každý svazek svůj záznam v katalogu a své čČNB, souborný záznam v tomto případě neexistuje
- popis nadřazené entity, kde tedy existuje pouze 1 katalogizační záznam pro více svazků monografie, nebude součástí metadat popisujících svazek

[294] Toto pořadí nevyjadřuje logickou strukturu dokumentu, ta je popsána jinde.

- ***ad svazek (Volume)*** – MODS popis svazku u klasické monografie (1 svazek = 1 záznam) odpovídá záznamu v katalogu
- ***ad vnitřní část (Internal Part)*** – bližší určení typů „textových oddílů“ a „obrazů“ (fotografie, tabulka, ilustrace, graf apod.) bude možné vyjádřit pomocí atributů a výrazů kontrolovaného slovníku v elementu <genre>
- ***ad příloha (Supplement)*** – přílohou se rozumí volně vložená entita do jednotlivého svazku, např. mapa, klíč (řešení úloh), pracovní sešit, CD/DVD apod. Rozlišujeme 3 druhy příloh monografie:
    1. příloha, která **se neskenuje**, ale chceme o ní vytvořit bibliografický záznam; dát najevo čtenáři, že existuje (např. CD/DVD apod.).
        - digitální podoba přílohy (pokud existuje) není součástí balíčku PSP svazku
        - popis lze udělat v rámci popisu přílohy (Supplement) v MODS – viz specifikace níže
        - pokud existuje záznam v katalogu k této příloze (např. CD/DVD, mapa apod.), bude využit pro generování MODS záznamu přílohy
        - taková příloha není součástí logické strukturální mapy standardu METS
    2. příloha podobného typu, tvaru a velikosti jako je popisovaný svazek monografie, která se spolu s číslem **skenuje.**
        - digitální podoba přílohy je, spolu se svazkem (Volume), součástí PSP balíčku svazku a je součástí hlavního METS záznamu
        - popis lze udělat v rámci popisu přílohy (Supplement) v MODS – viz specifikace níže
        - taková příloha může mít vnitřní části (InternalPart) stejně jako svazek (Volume) a jejich text je součástí ALTO XML, které je společné pro svazek (Volume) i přílohu (Supplement)
        - **taková příloha je součástí logické strukturální mapy standardu METS**
        - **taková příloha je součástí fyzické strukturální mapy standardu METS (linky mezi jednotlivými soubory reprezentujícími stránky a popisnými metadaty)**
    3. příloha odlišného typu typu, tvaru a velikosti než je popisovaný svazek monografie, která **se skenuje a popisuje zvlášť** na svazku nezávisle.
        - může jít např. o mapu apod.
        - k těmto přílohám vznikají metadata podobně jako pro jednotlivé svazky monografií, ovšem na původním svazku, ke kterému příloha patřila, nezávisle; tj. pro „původní“ svazek, u kterého byla příloha, vznikne 1 popis (PSP balíček s jedním hlavním METS záznamem a ALTO XML souborem) a pro přílohu je vytvořen další 1 popis (a PSP balíček s METS záznamem)

- v katalozích NK ČR a MZK neexistují údaje o kapitolách monografií – tj. vnitřní členění a popis musí vzniknout v digitalizaci, popis titulu/svazku monografie musí být stažen z katalogu do workflow digitalizace
- stránka se nebude popisovat, její logické i fyzické číslování i typ stránky je obsaženo ve struktuře METS dokumentu (část structMap)
    - typ stránky (Advertisement, Blank, Content, Index aj.) budou odpovídat přesně seznamu typů z DTD monografie – viz http://digit.nkp.cz/Monographs/DTD/2.10/Monograph.xsd

- pro každou entitu/úroveň záznamu vznikne jeden MODS záznam s vlastním ID, které bude označovat i typ části (např. oddíl, ilustrace apod.); v případě opakování částí se bude opakovat odpovídající počet MODS záznamů v jednom PSP balíčku
- každý MODS záznam entity bude uložen ve vlastní METS části <dmdSec> pomocí <mdWrap>
- u úrovní kde je to potřeba (vnitřní část, příloha apod.) se budou opakovat <dmdSec> části tolikrát, kolik je konkrétních částí
    - tj. v METS záznamu vznikne 1 část <dmdSec> pro bibliografický záznam svazku monografie, jedna nebo více <dmdSec> částí pro každou vnitřní části (pro všechny články i obrázky) a odpovídající počet <dmdSec> částí pro přílohy, dle počtu příloh
    - bibliografický popis obrazů bude velmi minimalistický
- **všechny top elementy MODS standardu jsou opakovatelné, kromě <recordInfo>**
- **všechny elementy Dublin Core jsou opakovatelné**

Každá část <dmdSec> musí mít ID (identifikátor) a vnořený element <mdWrap> s atributy MDTYPE, MIMETYPE.

element	atributy	popis	povinnost
<dmdSec>	ID	identifikátor <dmdSec> části METS záznamu ---------- ID: pro <dmdSec> s popisem svazku (titulu) monografie hodnota „MODSMD_VOLUME“ a „DCMD_VOLUME“ pro <dmdSec> s popisem vnitřní části monografie hodnota dle typů vnitřní části (oddíl[295], obraz) - hodnoty „MODSMD_CHAP“ a „DCMD_CHAP“ pro článek a hodnoty „MODSMD_PICT“ a „DCMD_PICT“ pro obraz pro <dmdSec> s popisem přílohy monografie hodnota „MODSMD_SUPPL“ a „DCMD_SUPPL“	M
<mdWrap>	MDTYPE MIMETYPE	element obsahující vložené záznamy MODS ---------- MDTYPE – hodnota „MODS“ pro záznamy v MODS, hodnota „DC“ pro záznam v Dublin Core MIMETYPE – hodnota „text/xml“	M

[295] Pozor výraz „kapitola“ je v tomto kontextu obecný a může vyjadřovat nejen kapitolu, ale také např. předmluvu, obsah apod.

8.4.1 Navrhovaná pole MODS a Dublin Core pro jednotlivé části monografie

8.4.1.1 Pole MODS a DC pro svazek monografie

Element MODS	Atributy	Popis	povinnost	Element DC
<titleInfo>	ID type	název svazku monografie pro plnění použít katalogizační záznam --- ID musí vyjadřovat název úrovně, tj. např. „MODS_VOLUME“ type: hodnota „alternative“ pro paralelní a jiné názvy (odpovídají poli 245 podpoli „b“)	M	
<title>		názvová informace – název svazku monografie hodnoty převzít z katalogu, odpovídá poli 245, podpoli „a“ pro hlavní název	M	<dc.title>
<subTitle>		podnázev svazku monografie	MA	<dc.title>
<partNumber>		číslo části, např. určité řady/edice (část 1, řada B)	R	<dc:description>
<partName>		jméno edice nebo speciální ediční řady, např. Knihy odvahy a dobrodružství	R	<dc:description>
<name>	type	údaje o odpovědnosti za svazek --- type: použít jeden z typů – personal – corporate – conference – family pokud má monografie autora a ilustrátora, element <name> se opakuje s různými rolemi POZOR – údaje o odpovědnosti nutno přebírat z polí 1XX a 7XX MARCu21	M	
<namePart>	type	údaje o křestním jméně a příjmení apod.	M	<dc:creator> nutno do

		nutno vyjádřit pro křestní jméno i příjmení ---------- type: použít jednu z hodnot: – date – doporučené pokud lze uvést – family – povinné pokud lze uvést – given – povinné pokud lze uvést – termsOfAddress – doporučené pokud lze uvést pokud nelze rozlišit křestní jméno a příjmení, nepoužije se atribut „type“ a jméno se zaznamená v podobě jaké je do jednoho elementu <namePart>		jednoho pole DC spojit jméno i příjmení
<role>		specifikace role osoby nebo organizace uvedené v elementu <name>	M	
<roleTerm>	type authority	popis role nutno použít kontrolovaný slovník např. z MARC21 ---------- type: code – kód role z kontrolovaného slovníku rolí http://www.loc.gov/marc/relators/relaterm.html authority – údaje o kontrolovaném slovníku využitém k popisu role, k popisu výše uvedeného MARC seznamu nutno uvést authority=“marcrelator“;	M	
<typeOfResource>		popis charakteristiky typu nebo obsahu zdroje jedna z hodnot: - text - cartographic - notated music - sound recording-musical - sound recording-	R	<dc:type>

		nonmusical - sound recording - still image - moving image - three dimensional object - software, multimedia - mixed material pro monografie hodnota text; mělo by se vyčítat z MARC21 katalogizačního záznamu z pozice 06 návěští		
<genre>		bližší údaje o typu dokumentu hodnota: volume	M	<dc:type>
<originInfo>		informace o původu předlohy Poznámka: Jeden nebo více výskytů elementů se předpokládá pro vydavatele, další výskyt v případě nutnosti popsat tiskaře. Pokud je nutno vyjádřit tiskaře (pole 260 podpole „f“ a „e“ a „g“ v MARC21), je nutno element <originInfo> opakovat s atributem transliteration=“printer“ a elementy <place>, <publisher>, <dateCreated>, které budou obsahovat údaje o tiskaři.	M	
<place>		údaje o místě spojeném s vydáním, výrobou nebo původem popisovaného dokumentu	MA	<dc:coverage>
<placeTerm>	type	konkrétní určení místa, např. Praha odpovídá hodnotě z katalogizačního záznamu, pole 260, podpole „a“ ---------- type – bude vždy text	MA	<dc:coverage>
<publisher>		jméno entity, která dokument	MA	<dc:publisher>

		vydala, vytiskla nebo jinak vyprodukovala odpovídá poli 260 podpoli „b“ katalogizačního záznamu v MARC21; Pokud má monografie více vydavatelů, přebírají se za záznamu všichni (jsou v jednom poli 260).		
<dateIssued>	qualifier	datum vydání předlohy, přebírat z katalogu; odpovídá hodnotě z katalogizačního záznamu, pole 260, podpole „c“ jiná data než rok možno zapsat v následujících podobách: - DD.MM.RRRR – pokud víme den, měsíc i rok vydání - MM.RRRR – pokud víme jen měsíc a rok vydání - RRRR – pokud víme pouze rok - DD.-DD.MM.RRRR – vydání pro více dní - MM.-MM.RRRR – vydání pro více měsíců --- qualifier – možnost dalšího upřesnění, hodnota „approximate“ pro data, kde nevíme přesný údaj	M	<dc:date>
<dateCreated>	qualifier	datum vytvoření předlohy bude použito pouze při popisu tiskaře, viz poznámka u elementu <originInfo> odpovídá hodnotě z katalogizačního záznamu, pole 260, podpole „g“ --- qualifier – možnost dalšího upřesnění, hodnota „approximate“ pro data, kde	R	

		nevíme přesný údaj		
<issuance>		údaje o vydávání hodnota monographic odpovídá hodnotě uvedené návěští MARC21 na pozici 07	M	
<language>		údaje o jazyce dokumentu; v případě vícenásobného výskytu nutno element <language> opakovat	M	
<languageTerm>	type authority objectPart	přesné určení jazyka – kódem nutno použít kontrolovaný slovník ISO 639-2, http://www.loc.gov/standards/iso639-2/php/code_list.php --- type: použít hodnotu code authority: použít hodnotu „iso639-2b“; odpovídá poli 041 MARC21, podpoli „a“ objectPart: možnost vyjádřit jazyk konkrétní části svazku; možné hodnoty např.: summary (pro shrnutí), original (pro předlohu u překladu) aj. – nutno vytvořit kontrolovaný slovník; jazyk resumé lze přebírat z pole 041, podpole „b“ jazyk předlohy u překladu lze přebírat z pole 041, podpole „h“	M	<dc:language>
<physicalDescription>		obsahuje údaje o fyzickém popisu zdroje/předlohy	M	
<form>	authority	údaje o fyzické podobě dokumentu, např. print, electronic apod. pro monografie hodnota print odpovídá hodnotám pozice 23 a 29 v poli 008 MARC21 --- authority: hodnota „marcform“	M	<dc:format>

<extent>		údaje o rozsahu (stran, svazků nebo rozměrů) odpovídá hodnotám v poli 300 podpolích „a“ a „c“ MARC21, pokud jsou vyplněna obě pole, bude se element <extent> opakovat; počet stránek bude vyjádřen ve fyzické strukturální mapě a bude tak vidět v aplikaci zpřístupnění i bez vyplnění tohoto pole	RA	<dc:format>
<note>		poznámka o fyzickém stavu dokumentu; pro každou poznámku je nutno vytvořit nový <note> element	RA	
<abstract>		shrnutí obsahu jako celku odpovídá poli 520 MARC21	R	<dc:description>
<note>		obecná poznámka ke svazku monografie jako celku odpovídá poli 500 v MARC21	RA	<dc:description>
<subject>	authority	údaje o věcném třídění předpokládá se přebírání z katalogizačního záznamu --- authority: vyplnit hodnotu „czenas“	R	
<topic>		libovolný výraz specifikující nebo charakterizující obsah svazku monografie; použít kontrolovaný slovník - např. z báze autorit AUT NK ČR (věcné téma) nebo obsah pole 650 záznamu MARC21	M	<dc:subject>
<geographic>		geografické věcné třídění použít kontrolovaný slovník - např. z báze autorit AUT NK ČR (geografický termín) nebo obsah pole 651 záznamu MARC21	R	<dc:subject>
<temporal>		chronologické věcné třídění použít kontrolovaný slovník - např. z báze autorit AUT NK ČR (chronologický údaj) nebo obsah	R	<dc:subject>

		pole 648 záznamu MARC21		
<name>		jméno použité jako věcné záhlaví použít kontrolovaný slovník - např. z báze autorit AUT NK ČR (jméno osobní) nebo obsah pole 600 záznamu MARC21	R	<dc:subject>
<classification>	authority	klasifikační údaje věcného třídění podle Mezinárodního desetinného třídění odpovídá poli 080 MARC21 ---------- authority: vyplnit hodnotu „udc“	M	<dc:subject>
<relatedItem>	type	informace o dalších dokumentech/částech/zdrojích, které jsou ve vztahu k popisovanému dokumentu; Poznámka: element <relatedItem> může obsahovat jakýkoliv jiný element MODS – jejich použití se řídí pravidly popsanými pro tyto elementy; ---------- type: hodnota „series“	RA	
<identifier>	type	údaje o identifikátorech, obsahuje unikátní identifikátory mezinárodní nebo lokální, které svazek monografie má – viz přehled typů atributů níže ---------- type: budou se povinně vyplňovat následující hodnoty, pokud existují: - doi - hdl - handle - issn - převzít z katalogizačního záznamu NK ČR - isbn - převzít z katalogizačního záznamu NK ČR - ccnb – čČNB - převzít z katalogizačního záznamu NK ČR	M	<dc:identifier>

		- permalink záznamu z katalogu NK ČR, např. http://aleph.nkp.cz/F/?func=direct&doc_number=002186258&local_base=NKC - urnnbn - pro URN:NBN, např. zápis ve tvaru urn:nbn:cz:ndk-123456 pro projekt NDK; pozor, musí odpovídat URN:NBN, podle kterého je pojmenovaný PSP balíček a jeho jednotlivé soubory - uuid - jiný interní identifikátor, hodnota atributu „local“, lze použít např. k vyjádření čárového kódu		
<location>		údaje o uložení popisovaného dokumentu, např. signatura, místo uložení apod.	MA	
<url>	note	pro uvedení lokace elektronického dokumentu --- note: pro poznámku o typu URL (na plný text, abstrakt apod.)	O	<dc:source>
<physicalLocation>	authority	údaje o instituci, kde je fyzicky uložen popisovaný dokument, např. NK ČR nutno použít kontrolovaný slovník – sigly knihovnen (ABA001 atd.) odpovídá poli 040 v MARC21 pozn. u dokumentů v digitální podobě není možné vyplnit --- authority: hodnota „siglaADR“	M	<dc:source>
<shelfLocator>		signatura nebo lokační údaje o dokumentu	M	<dc:source>

<part>	type	popis části, pokud je svazek části souboru, element může být využit jen na zaznamenání <caption> --- type: hodnota bude vždy „volume“	O	
<caption>		text před označením čísla, např. „č.“, „část“, „No.“ apod.	RA	
<recordInfo>		údaje o metadatovém záznamu – jeho vzniku, změnách apod.	M	
<recordContentSource>		kód nebo jméno instituce, která záznam vytvořila nebo změnila; nutno vytvořit kontrolovaný slovník	R	
<recordCreationDate>	encoding	datum prvního vytvoření záznamu, na úroveň minut --- encoding: záznam bude podle normy ISO 8601 na úroveň minut, hodnota atributu tedy iso8601	M	
<recordChangeDate>	encoding	datum změny záznamu --- encoding: záznam bude podle normy ISO 8601 na úroveň minut, hodnota atributu tedy iso8601	R	
<recordOrigin>		údaje o vzniku záznamu hodnoty: machine generated nebo human prepared	R	

8.4.1.2 Pole MODS a DC pro vnitřní část monografie (textový oddíl a obraz)

Element MODS	Atributy	Popis	Povinnost	Element DC
<titleInfo>	ID	názvová informace vnitřní části --- ID musí vyjadřovat název úrovně, tj. např. „MODS_PICTURE“ pro obrázek v textu, „MODS_CHAPTER“ pro textový oddíl apod.	M	
<title>		vlastní název vnitřní části (oddílu, obrazu); u obrazu brát případně z popisku	M	<dc:title>

		obrazu; pokud není titul, nutno vyplnit hodnotu „untitled“		
<subTitle>		podnázev vnitřní části (oddílu); např. podnázev kapitoly	MA	<dc:title>
<partNumber>		číslo vnitřní části	RA	<dc:title>
<partName>		název vnitřní části	RA	<dc:title>
<name>	type	údaje o odpovědnosti za vnitřní část (oddíl i obraz) --- type: použít jeden z typů – personal – corporate – conference – family	MA	
<namePart>	type	údaje o křestním jméně a příjmení apod. nutno vyjádřit pro křestní jméno i příjmení --- type: použít jednu z hodnot: – date – doporučené pokud lze uvést – family – povinné pokud lze uvést – given – povinné pokud lze uvést – termsOfAddress – doporučené pokud lze uvést pokud nelze rozlišit křestní jméno a příjmení, nepoužije se atribut „type“ a jméno se zaznamená v podobě jaké je do jednoho elementu <namePart>	MA	<dc:creator> nutno do jednoho pole DC spojit jméno i příjmení
<role>		specifikace role osoby nebo organizace uvedené v elementu <name>	MA	
<roleTerm>	type authority	popis role nutno použít kontrolovaný slovník např. z MARC21	MA	

		---------- type: code – kód role z kontrolovaného slovníku rolí http://www.loc.gov/marc/relators/relaterm.html) authority – údaje o kontrolovaném slovníku využitém k popisu role, k popisu výše uvedeného MARC seznamu nutno uvést authority="marcrelator"		
<genre>	type	bližší údaje o typu vnitřní části povinné hodnota: chapter nebo picture ---------- type: doporučené hodnota pro chapter – možnost vyplnit bližší určení typu oddílu (možnost použít DTD monografie, MonographComponentPart Types) - table of content - advertisement - abstract - introduction - review - dedication - bibliography - editorsNote - preface - chapter - article - index (použije se pro všechny typy seznamů mimo hlavní obsah; např. seznam obrazů, tabulek apod.) - unspecified – pokud nepatří ani do jedné z výše uvedených kategorií - aj. hodnota pro picture – možnost vyplnit další určení typu obrazu	M	<dc:type>

		- table - illustration - chart - photograph - graphic - map - advertisement - cover - unspecified – pokud nepatří ani do jedné z výše uvedených kategorií - aj.		
<language>		údaje o jazyce vnitřní části nelze plnit u obrazu; v případě vícenásobného výskytu nutno element <language> opakovat	MA	
<languageTerm>	type authority	přesné určení jazyka – kódem nutno použít kontrolovaný slovník ISO 639-2, http://www.loc.gov/standards/iso639-2/php/code_list.php nelze plnit u obrazu --- type: použít hodnotu code authority: použít hodnotu „iso639-2b“	M	<dc:language>
<physicalDescription>		obsahuje údaje o fyzickém popisu vnitřní části; určeno spíše pro oddíly než pro obrazy	R	
<form>	authority	údaje o fyzické podobě vnitřní části, např. print, electronic apod. --- authority: hodnota „marcform“	R	<dc:format>
<abstract>		shrnutí obsahu vnitřní části	R	<dc:description>
<note>		obecná poznámka k vnitřní části do poznámky by se měla dávat šifra autora vnitřní části, která se vyskytuje pod vnitřní částí	RA	<dc:description>
<subject>		údaje o věcném třídění	R	

<topic>	authority (volitelné)	libovolný výraz specifikující nebo charakterizující obsah vnitřní části; lze (není ovšem nutno) použít kontrolovaný slovník - např. z báze autorit AUT NK ČR (věcné téma) --- při použití autoritních záznamů použít AUT NK ČR a atribut authority: vyplnit hodnotu „czenas“; při použití volných klíčových slov atribut authority nepoužívat	M	<dc:subject>
<geographic>	authority	geografické věcné třídění použít kontrolovaný slovník - např. z báze autorit AUT NK ČR (geografický termín) --- authority: vyplnit hodnotu „czenas“	R	<dc:subject>
<temporal>	authority	chronologické věcné třídění použít kontrolovaný slovník - např. z báze autorit AUT NK ČR (chronologický údaj) --- authority: vyplnit hodnotu „czenas“	R	<dc:subject>
<name>	authority	jméno použité jako věcné záhlaví použít kontrolovaný slovník - např. z báze autorit AUT NK ČR (jméno osobní) --- authority: vyplnit hodnotu „czenas“	R	<dc:subject>
<classification>	authority	klasifikační údaje věcného třídění podle Mezinárodního desetinného třídění plnit pouze pro oddíl odpovídá poli 080 MARC21 --- authority: vyplnit hodnotu „udc“	RA	<dc:subject>
<identifier>	type	údaje o identifikátorech, obsahuje unikátní identifikátory mezinárodní nebo lokální, které vnitřní část má	M	<dc:identifier> povinné

		– viz přehled typů atributů níže ---------- type: budou se povinně vyplňovat následující hodnoty, pokud existují pro oddíl nebo obraz: - doi - hdl - handle - urnnbn - pro URN:NBN, u vnitřních částí monografií se s URN:NBN počítá primárně pro články ve sborníku, ne pro „obyčejné“ kapitoly - uuid - jiný interní identifikátor, hodnota atributu „local“, lze použít např. k vyjádření čárového kódu		
<part>		vrchní element, který bude použit pouze na záznam rozsahu vnitřní části; nelze u obrazu	RA	
<extent>		upřesnění popisu části – rozsah na stránkách	MA	<dc:format>
<start>		první stránka, na které vnitřní část začíná	MA	<dc:coverage>
<end>		poslední stránka, na které vnitřní část končí	MA	<dc:coverage>
<recordInfo>		údaje o metadatovém záznamu vnitřní části – jeho vzniku, změnách apod.	M	
<recordContentSource>		kód nebo jméno instituce, která záznam vytvořila nebo změnila; nutno vytvořit kontrolovaný slovník	R	
<recordCreationDate>	encoding	datum prvního vytvoření záznamu vnitřní části ---------- encoding: záznam bude podle normy ISO 8601 na úroveň minut, hodnota atributu tedy iso8601	M	
<recordChangeDate>	encoding	datum změny záznamu vnitřní části ---------- encoding: záznam bude podle normy ISO 8601 na úroveň minut,	R	

		hodnota atributu tedy iso8601		
<recordOrigin>		údaje o vzniku záznamu vnitřní části hodnoty: machine generated nebo human prepared	R	

8.4.1.3 Pole MODS a DC pro přílohu

Element MODS	Atributy	Popis	Povinnost	Element DC
<titleInfo>	ID	názvová informace přílohy použít názvové autority nebo katalogizační záznam --- ID musí vyjadřovat název úrovně, tj. „MODS_SUPPLEMENT“	M	
<title>		názvová informace – název svazku monografie, jehož součástí příloha je; převzít z katalogu	M	<dc:title>
<partNumber>		číslo přílohy, pokud nějaké má doporučené pokud lze vyplnit	MA	<dc:description>
<partName>		název přílohy	MA	<dc:title>
<name>	type	údaje o odpovědnosti za přílohu --- type: použít jeden z typů – personal – corporate – conference – family	MA	
<namePart>	type	údaje o křestním jméně a příjmení apod. nutno vyjádřit pro křestní jméno i příjmení --- type: použít jednu z hodnot: – date – doporučené pokud lze uvést – family – povinné pokud lze uvést – given – povinné pokud lze uvést – termsOfAddress –	MA	<dc:creator> nutno do jednoho pole DC spojit jméno i příjmení

		doporučené pokud lze uvést pokud nelze rozlišit křestní jméno a příjmení, nepoužije se atribut „type“ a jméno se zaznamená v podobě jaké je do jednoho elementu <namePart>		
<role>		specifikace role osoby nebo organizace uvedené v elementu <name>	MA	
<roleTerm>	type authority	popis role nutno použít kontrolovaný slovník např. z MARC21 --- type: code – kód role z kontrolovaného slovníku rolí http://www.loc.gov/marc/relators/relaterm.html) authority – údaje o kontrolovaném slovníku využitém k popisu role, k popisu výše uvedeného MARC seznamu nutno uvést authority=“marcrelator“	MA	
<typeOfResource>		popis charakteristiky typu nebo obsahu přílohy jedna z hodnot: - text – např. pro přílohu typu časopis, kniha, brožura apod. - cartographic – pro mapy - notated music - sound recording-musical - pro hudební CD/DVD - sound recording-nonmusical - sound recording - still image – fotografie, plakáty apod. - moving image – pro filmová DVD - three dimensional object - software, multimedia – pro	R	<dc:type>

		CD/DVD se SW - mixed material		
<genre>		bližší údaje o typu dokumentu hodnota: supplement	M	<dc:type>
<originInfo>		informace o původu přílohy *plnit pokud se liší od údajů v popisu svazku monografie (platí i pro jednotlivé sub-elementy)* Poznámka: Jeden nebo více výskytů elementů se předpokládá pro vydavatele, další výskyt v případě nutnosti popsat tiskaře. Pokud je nutno vyjádřit tiskaře (pole 260 podpole „f“ a „e“ a „g“ v MARC21), je nutno element <originInfo> opakovat s atributem transliteration=“printer“ a elementy <place>, <publisher>, <dateCreated>, které budou obsahovat údaje o tiskaři.	MA	
<place>		údaje o místě spojeném s vydáním, výrobou nebo původem přílohy	MA	<dc:coverage>
<placeTerm>	type	konkrétní určení místa, např. Praha odpovídá hodnotě katalogizačního záznamu, pole 260, podpole „a“ --- type – bude vždy text	MA	<dc:coverage>
<publisher>		jméno entity, která přílohu vydala, vytiskla nebo jinak vyprodukovala odpovídá poli 260 podpoli „b“ katalogizačního záznamu v MARC21	MA	<dc:publisher>
<dateIssued>	qualifier	datum vydání přílohy, dle toho jaké údaje jsou k dispozici jiná data než rok možno zapsat v následujících podobách: - DD.MM.RRRR – pokud víme den, měsíc i rok vydání	MA	<dc:date>

		- MM.RRRR – pokud víme jen měsíc a rok vydání - RRRR – pokud víme pouze rok - DD.-DD.MM.RRRR – vydání pro více dní - MM.-MM.RRRR – vydání pro více měsíců možno použít hodnotu z katalogizačního záznamu, pole 260, podpole „c“ --- qualifier – možnost dalšího upřesnění, hodnota „approximate“ pro data, kde nevíme přesný údaj		
<dateCreated>	qualifier	datum vytvoření přílohy bude použito pouze při popisu tiskaře, viz poznámka u elementu <originInfo> nebo např. u popisu CD/DVD apod. odpovídá hodnotě z katalogizačního záznamu, pole 260, podpole „g“ --- qualifier – možnost dalšího upřesnění, hodnota „approximate“ pro data, kde nevíme přesný údaj	R	
<frequency>		údaje o pravidelnosti vydávání odpovídá údaji MARC21 v poli 310 nebo pozici 18 v poli 008	RA	
<language>		údaje o jazyce dokumentu	M	
<languageTerm>	type authority	přesné určení jazyka – kódem nutno použít kontrolovaný slovník ISO 639-2, http://www.loc.gov/standards/iso639-2/php/code_list.php --- type: použít hodnotu code authority: použít hodnotu „iso639-2b“	M	<dc:language>

<physicalDescription>		obsahuje údaje o fyzickém popisu	M	
<form>	authority	údaje o fyzické podobě dokumentu, např. print, electronic apod. povinné pro tištěné předlohy hodnota „print“, pro elektronické přílohy „electronic“ odpovídá hodnotám pozice 23 a 29 v poli 008 MARC21 --- authority: hodnota „marcform“	M	<dc:format>
<extent>		údaje o rozsahu (stran, svazků nebo rozměrů) odpovídá hodnotám v poli 300 podpolích „a“ a „c“ MARC21, pokud jsou vyplněna obě pole, bude se element <extent> opakovat	RA	<dc:format>
<note>		poznámka o fyzickém stavu dokumentu; pro každou poznámku je nutno vytvořit nový <note> element	RA	
<abstract>		shrnutí obsahu dokumentu odpovídá poli 520 MARC21	RA	<dc:description>
<note>		obecná poznámka k dokumentu odpovídá poli 500 v MARC21	RA	<dc:description>
<subject>	authority	údaje o věcném třídění --- authority: vyplnit hodnotu „czenas“	R	
<topic>		libovolný výraz specifikující nebo charakterizující obsah přílohy; použít kontrolovaný slovník - např. z báze autorit AUT NK ČR (věcné téma)	M	<dc:subject>
<geographic>		geografické věcné třídění použít kontrolovaný slovník - např. z báze autorit AUT NK ČR (geografický termín)	R	<dc:subject>
<temporal>		chronologické věcné třídění použít kontrolovaný slovník - např. z báze autorit AUT NK ČR	R	<dc:subject>

		(chronologický údaj)		
<name>		jméno použité jako věcné záhlaví použít kontrolovaný slovník - např. z báze autorit AUT NK ČR (jméno osobní)	R	<dc:subject>
<classification>	authority	klasifikační údaje věcného třídění podle Mezinárodního desetinného třídění odpovídá poli 080 MARC21 ---------- authority: vyplnit hodnotu „udc“	M	<dc:subject>
<identifier>	type	údaje o identifikátorech, obsahuje unikátní identifikátory mezinárodní nebo lokální, které příloha má – viz přehled typů atributů níže ---------- type: budou se povinně vyplňovat následující hodnoty, pokud existují: - doi - hdl - handle - issn - převzít z katalogizačního záznam NK ČR - isbn - převzít z katalogizačního záznam NK ČR - ccnb – čČNB - převzít z katalogizačního záznam NK ČR - permalink záznamu z katalogu NK ČR, např. http://aleph.nkp.cz/F/?func=direct&doc_number=002186258&local_base=NKC - urnnbn - pro URN:NBN - uuid - jiný interní identifikátor, hodnota atributu „local“, lze použít např. k vyjádření čárového kódu	MA	<dc:identifier>

8.5 METS část <amdSec> - Technická a administrativní metadata (MIX a PREMIS)

- technická metadata jsou určena primárně pro zachycení technických informací o formátech souborů, o výsledcích validací a kontrol
- administrativní metadata zachycují veškeré změny, procesy apod., které byly na datech i metadatech provedeny
- pro všechny typy digitálních objektů se bude využívat standard PREMIS (jeho části Object, Event a Agent); pro obrazová data pak navíc také standard MIX
- technická a administrativní metadata budou vznikat i pro prvotní sken (většinou TIFF), který se po nutných úpravách maže a dále neuchovává, metadata se ovšem uchovávají
- technická a administrativní metadata pro různé reprezentace jedné strany svazku monografie (původní TIFF, MC, ALTO XML a OCR.TXT) budou zabalena v části <amdSec> vedlejšího METS záznamu (AMD_METS.xml) ve vlastních schématech (MIX, PREMIS – části Object; Events; Agent)
    - vedlejší METS záznam (AMD_METS.xml) a je linkován z hlavního METS záznamu dokumentu
- **pro všechny reprezentace jedné strany svazku monografie bude ve vedlejším METS záznamu AMD_METS.xml existovat jedna část <amdSec>, která bude obsahovat metadata v <techMD> a <digiprovMD> podčástech pro jednotlivé soubory**
- **plnění technických metadat se předpokládá z výstupů vzniklých využitím služeb třetích stran, jako jsou JHOVE2, DROID aj.**

Část <amdSec> musí mít atribut ID a vnořený element <techMD> nebo <digiprovMD>, oba s atributem ID a vnořeným elementem <mdWrap> s atributem MDTYPE.

element	atributy	popis	Povinnost
<amdSec>	ID	element obsahující technická metadata ve standardu PREMIS nebo MIX ---------- ID – identifikátor konkrétní části <amdSec>, např. pro stránku 1 by hodnota mohla být „PAGE0001“	M
<techMD> nebo <digiprovMD>	ID	element rozlišující typy jednotlivých administrativních metadat ---------- ID pro část <techMD>: – pro části obsahující PREMIS-object hodnota „OBJ_001“ – objekt 1 (PREMIS Object pro smazaný TIFF, OBJ_002 by bylo pro MC, OBJ_003 pro ALTO XML – pro části obsahující MIX hodnota „MIX_001“ = MIX metadata pro původní TIFF, „MIX_002“ pro MC pro část <digiprovMD>:	M

		– pro části obsahující PREMIS-event hodnota „EVT_001“ apod. – pro části obsahující PREMIS-agent hodnota „AGENT_001“ apod.	
<mdWrap>	MDTYPE	element obsahující vložené záznamy PREMIS, MIX ---------- MDTYPE – pro záznamy PREMIS Object, event i agent vždy hodnota „PREMIS“ – pro záznamy MIX hodnota „NISOIMG“	M

8.5.1 PREMIS Objects

- bude odpovídat poslední aktuální verzi v době implementace (leden 2011 - PREMIS data dictionary v. 2.1), nebo verzi předchozí
- popisovat se pomocí PREMIS Object budou soubory, tj. dle specifikace PREMIS vždy úroveň tzv. File (ne reprezentace ani bitstream)
- záznam v PREMIS Object se bude vytvářet pro následující soubory vzniklé v procesu digitalizace: 1) původní sken; 2) archivní obrazové kopie; 3) ALTO XML
- PREMIS Object se nebude vytvářet pro OCR.TXT soubory
- pro každý záznam PREMIS Object bude existovat vlastní podčást <techMD>
- záznam PREMIS Object pro jeden soubor bude obsahovat linky na Události, které jsou popsány v PREMIS Events ve stejném METS metadatovém záznamu konkrétního dokumentu (svazku monografie) v části <digiprovMD>; vazba bude provedena přes element <premis:relatedEventIdentification>; to samé platí pro objekty, které budou nalinkovány v případě vztahu (např. UC vznikla z MC) s popisovaným objektem přes element <premis:relatedObjectIdentification>
    - např. PREMIS Object záznam popisující archivní soubor JPEG 2000 je tímto způsobem nalinkován na původní sken ve formátu TIFF (resp. na jeho PREMIS Object záznam) pomocí elementu <relatedObjectIdentification>, který obsahuje ID původního objektu (např. TIFF)
    - zároveň pomocí elementu <relatedEventIdentification> je záznam PREMIS Object archivního souboru JPEG 2000 nalinkován na událost, během které vznikl
- **POZOR – PREMIS Object bude vznikat a uchovávat se i pro neexistující data (původní a posléze smazaný TIFF)**

Pole záznamu PREMIS Object

Element	Popis	Povinnost	Použití pro
<objectIdentifier>	identifikátor k jednoznačnému odlišení objektu v určitém kontextu; 1-n	M	MC, XML, PS
<objectIdentifierType>	popis kontextu, ve kterém je identifikátor unikátní, např. NDK, ANL nebo název repozitáře;	M	MC, XML,

	nutno použít kontrolovaný slovník; 1-1		PS
<objectIdentifierValue>	vlastní hodnota identifikátoru, např. img0001-master, urn.nbn.cz-123465 apod.; 1-1	M	MC, XML, PS
<objectCategory>	typ objektu, ke kterým se metadata (PREMIS Object) vztahuje, např. file pro soubor, representation pro dig. reprezentaci, bitstream pro bitstream; 1-1	M	MC, XML, PS
<preservationLevel>	údaje o úrovni ochrany souboru, která se na něj vztahuje; některé soubory nejsou tak důležité jako jiné, mají menší úroveň ochrany; 0-n	M	MC, XML, PS
<preservationLevelValue>	hodnota úrovně ochrany, která je pro soubor relevantní, pro původní sken PS hodnota deleted, pro MC a XML hodnota preservation; 1-1	M	MC, XML, PS
<preservationLevelDateAssigned>	datum, kdy byla přiřazena hodnota úrovně ochrany, zápis v ISO 8601, na úroveň dne (DD-MM-RRRR) 0-1	R	MC, XML, PS
<objectCharacteristics>	technické údaje o souboru 1-n	M	MC, XML, PS
<compositionLevel>	údaj o tom, zda je nutné digitální objekt rozbalit nebo dekódovat; např. 0 (defaultně pro žádné zabalení nebo kódování); 1 pro jedno zabalení a kódovaní, podobně pak hodnota 2; 1-1	M	MC, XML, PS
<fixity>	údaje o kontrolním součtu 0-n	M	MC, XML, PS
<messageDigestAlgorithm>	použitý algoritmus kontrolního součtu, např. MD5 aj. 1-1	M	MC, XML, PS
<messageDigest>	hodnota kontrolního součtu 1-1	M	MC, XML, PS
<messageDigestOriginator>	agent (osoba, instituce, stroj, SW), který kontrolní součet vytvořil (např. JHOVE apod.) 0-1	M	MC, XML, PS
<size>	údaje o velikosti souboru v bytech 0-1	M	MC, XML,

			PS
<format>	údaje o formátu souboru 1-n	M	MC, XML, PS
<formatDesignation>	identifikace formátu souboru, výstup z JHOVE, PRONOM služeb apod. 0-1	M	MC, XML, PS
<formatName>	jméno formátu, např. image/tiff nebo Adobe PDF 1-1	M	MC, XML, PS
<formatVersion>	verze formátu, např. 6.0 0-1	M	MC, XML, PS
<formatRegistry>	identifikace formátu – dodatečná informace o záznamu formátů v registrech formátů (např. PRONOM aj.) 0-1	M	MC, XML, PS
<formatRegistryName>	jméno použitého registru formátů, např. UDFR, PRONOM aj. 1-1	M	MC, XML, PS
<formatRegistryKey>	unikátní identifikátor (označení) formátu v registru, např. fmt/155 z PRONOM 1-1	M	MC, XML, PS
<creatingApplication>	údaje o aplikaci, ve které byl popisovaný soubor vytvořen; nutno popsat skener, SW kde vzniklo ALTO XML/TXT, SW/kodek pro vytvoření JPEG 2000 MC 0-n	M	MC, XML, PS
<creatingApplicationName>	název aplikace, např. ImageGear, Kakadu apod.; 0-1	M	MC, XML, PS
<creatingApplicationVersion>	verze aplikace, např. 15.03.000 0-1	M	MC, XML, PS
<dateCreatedByApplication>	datum a čas vytvoření, např. 2008-11-10T12:37:46; musí být ve tvaru ISO 8601 (na úroveň vteřin); 0-1	M	MC, XML, PS
<originalName>	původní jméno souboru, např. digibok_2007081301091_0011.jp2 0-1	M	MC, XML, PS
<relationship>	vyjádření vztahu popisovaného souboru k jiným souborům a událostem (eventům) 0-n	M	MC, XML

<relationshipType>	typ vztahu, doporučené hodnoty: derivation= vztah kde objekt je výsledkem změny jiného objektu; structural= vztah mezi částmi objektu; tj. např. ALTO vytvořené z TIFFU bude mít vztah derivation, podobně jako JPEG 2000 z TIFFu vytvořený; 1-1	M	MC, XML;
<relationshipSubType>	upřesnění vztahu, doporučené hodnoty: created from; has source; is source of; has sibling; has part; is part of; has root; includes; is included in; apod.; tj. např. ALTO nebo JPEG 2000 vytvořený z původního TIFFu budou mít vztah „created from“ 1-1	M	MC, XML;
<relatedObjectIdentification>	identifikace souvisejícího souboru 1-n pro MC, XML pro vyjádření vztahu k původnímu objektu (skenu)	M	MC, XML
<relatedObjectIdentifierType>	specifikace kontextu, ve kterém je identifikátor souboru jedinečný, např. URN; temporary filepath; objectID 1-1	M	MC, XML
<relatedObjectIdentifierValue>	vlastní řetězec identifikátoru, např. URN:NBN:cz-1301091_011#0001 nebo název souboru, cesta k souboru apod. 1-1	M	MC, XML
<relatedEventIdentification>	identifikace s popisovaným souborem související události (eventu); seznam událostí viz PREMIS Event 0-n	M	MC, XML
<relatedEventIdentifierType>	typ události, např. interní číslovací systém událostí jako no.nb.evt; NK repository event ID, UUID apod. 1-1	M	MC, XML
<relatedEventIdentifierValue>	hodnota identifikátoru události, např. NK_EVT_005 nebo hodnota UUID aj. 1-1	M	MC, XML
<relatedEventSequence>	pořadí události, např. 003; k určení pořadí lze určit datum události 0-1	R	MC, XML
<linkingEventIdentifier>	identifikátor události týkající původního skenu PS; typy událostí mohou být např. vytvoření, smazání	M	PS

	0-n pro PS nutný link na události vytvoření (digitalizace) a jeho vymazání		
<linkingEventIdentifierType>	typ identifikátoru události, např. UUID, NK_eventID, vlastní číslovací systém apod. 1-1	M	PS
<linkingEventIdentifierValue>	hodnota identifikátoru, např. event_01; img0001-master-event001 apod. 1-1	M	PS

8.5.2 PREMIS Event

- bude odpovídat poslední aktuální verzi v době implementace (leden 2011 - PREMIS data dictionary v. 2.1), nebo verzi předchozí
- PREMIS Event záznamy shromažďují informace o procesech a událostech, které se týkají jednoho nebo více objektů, v našem případě souborů; primární použití je k zaznamenání událostí, které popisovaný soubor mění nebo upravují.
- bude vznikat pro události, které se prováděly na obrazových datech
    - digitalizace – vytvoření prvního skenu (např. do TIFF)
    - vytvoření ALTO XML
    - vygenerování MC
    - vygenerování UC
    - vymazání PS
- popis událostí bude zachycovat informace o jejich výsledku/výstupu
- záznamy PREMIS Event budou uloženy v METS záznamu určeném pro administrativní a technická metadata (AMD_METS.xml) v jeho části <amdSec>, podčást <digiprovMD>
- pro každou událost bude vytvořena jedna <digiprovMD> část
- každý záznam PREMIS Event je linkován na původce aktivity, tedy na PREMIS Agent záznam

Pole záznamu PREMIS Event

Element	Popis	Povinnost
<eventIdentifier>	údaje o identifikátoru události v kontextu digitalizace nebo repozitáře 1-1	M
<eventIdentifierType>	typ identifikátoru, např. no.nb.evt; NK_eventID, UUID apod. 1-1	M
<eventIdentifierValue>	hodnota identifikátoru, např. EVT_001; event_019 apod. 1-1	M
<eventType>	kategorizace události, nutno použít kontrolovaný slovník; typy událostí, které musí být zaznamenány: capture, migration, derivation, deletion	M

	1-1	
<eventDateTime>	datum a čas kdy byla událost provedena; nutno zapsat v ISO 8601 na úroveň vteřin 1-1	M
<eventDetail>	další údaje o události, doporučené hodnoty pro výše uvedené <eventType> následují za /: – capture/digitization – vznik prvního skenu – capture/XML_creation – capture/TXT_creation – migration/MC_creation – derivation/UC_creation – deletion/PS_deletion 0-1	M
<eventOutcomeInformation>	informace o výsledku události 0-n	R
<eventOutcome>	kategorizace výsledku události, např. slovy jako successful nebo failure, možno použít kódy – nutno používat kontrolovaný slovník nebo seznam kódů 0-1	M
<linkingAgentIdentifier>	identifikace jednoho nebo více agentů spojených s událostí 0-n	M
<linkingAgentIdentifierType>	označení typu identifikátoru, např. NK_AgentID, UUID apod. 1-1	M
<linkingAgentIdentifierValue>	hodnota identifikátoru, např. agent_softwareName_5.2; agent_novakJ apod. 1-1	M
<linkingAgentRole>	role agenta ve vztahu k události, např. software; SW component; operator; nutno používat kontrolovaný slovník 0-n	R
<linkingObjectIdentifier>	informace o objektu/souboru spojeného s událostí, link na něj 0-n	M
<linkingObjectIdentifierType>	označení typu identifikátoru, např. PhysUnitID; URN, NK_OBJ, OBJ_001 apod.; hodnoty by se měly brát z kontrolovaného slovníku 1-1	M
<linkingObjectIdentifierValue>	hodnota identifikátoru, např. URN:NBN:cz-_0011#0001 aj. 1-1	M

8.5.3 PREMIS Agent

- bude odpovídat poslední aktuální verzi v době implementace (leden 2011 - PREMIS data dictionary v. 2.1), nebo verzi předchozí
- záznam PREMIS Agent obsahuje charakteristiku tzv. agenta, který je spojen s provedenou a zaznamenanou událostí (PREMIS Event)
    - agent může být osoba, organizace nebo software
- z PREMIS Event je linkováno na agenta, který určitou akci provedl, typ ID agenta a jeho hodnota jsou uvedené v PREMIS Events (<premis:linkingAgentIdentifier>), plný popis agenta je pak v PREMIS Agent
- záznamy PREMIS Agent budou uloženy v METS záznamu určeném pro administrativní a technická metadata (AMD_METS.xml) v jeho části <amdSec>, podčást <digiprovMD>
- pro každého agenta, tj. jeden PREMIS Agent záznam, bude vytvořena jedna <digiprovMD> část
- **informace v PREMIS Event a PREMIS Object přicházející z procesu digitalizace v PSP balíčku jsou dostačující a dají nám dostatečné informace o událostech, které se odehrály v souvislosti se vznikem digitálního objektu**
    - **další upřesnění události v PREMIS Agent není nutné**

Pole záznamu PREMIS Agent

Element	Popis	Povinnost
<agentIdentifier>	popis identifikátoru, který jednoznačně označuje agenta v rámci jednoho kontextu (repozitář např.) 1-n	M
<agentIdentifierType>	označení typu identifikátoru, např. NK_AgentID, UUID apod. 1-1	M
<agentIdentifierValue>	hodnota identifikátoru, např. agent_softwareName_5.2; agent_novakJ apod. 1-1	M
<agentName>	textové upřesnění agenta, např. přesný název SW, plné jméno osoby apod. - FixImage1.3; Jan Novák; CCS docWorks 6.2.1; 0-n	R
<agentType>	obecné označení agenta – pro osoby např. osoba, pro SW např. software apod. hodnoty: organization; person; software 0-1	M
<agentNote>	použití pouze pokud je <agentType> Software a půjde o agenta souvisejícího s migrací TIFF na JPEG 2000 (creation/migration Event); bude obsahovat příkaz k výrobě JPEG 2000 souboru v aplikaci Kakadu 0-n	MA

8.5.4 Technická metadata MIX

- bude využit standard MIX, verze aktuální v době implementace projektu, nebo verze předchozí (prosinec 2010 verze 2 – viz http://www.loc.gov/standards/mix//)
- **MIX záznam vzniká pouze pro obrazové soubory!**
  - **tj. bude vznikat 1) jeden záznam pro archivní kopii, 2) další záznam pro původní soubor vzniklý prvotním skenováním (nejčastěji TIFF**) a to i přesto, že tento TIFF se v průběhu výroby maže a není archivován
  - tyto dva MIX záznamy budou součástí jednoho METS záznamu AMD_METS.xml (v části <amdSec>, podčást <techMD>) pro administrativní a technická metadata, který vznikne ke každému obrazovému souboru a který je linkován z hlavního METS záznamu svazku monografie
- **MIX záznamy jednotlivých obrazových souborů se budou lišit – MIX záznam původního skenu nebude obsahovat např. element <ImageProcessing>, MIX záznam archivního souboru MC nebude naproti tomu obsahovat informace o procesu skenování, které se váží k původnímu skenu a budou v elementu <ImageCaptureMetadata> apod. – podrobnosti viz tabulka níže, sloupec „užití pro MC a PS“**
- **pro každý záznam MIX bude vytvořena vlastní část <techMD>**
- MIX může být také zapouzdřen v PREMIS Object <premis:objectCharacteristicsExtension>
- **externí služby, jako např. JHOVE a DROID, budou využívány k plnění polí standardu MIX**
- ve standardu MIX nebude uvedena informace o kontrolních součtech (fixity), která je obsažena v PREMIS Object a není nutno ji opakovat (viz MIX profily Nizozemí, Finska a Norska)
- element <fileSize> je pouze doporučený, údaj o velikosti souboru je součástí popisu PREMIS Object

Pole standardu MIX pro popis archivní kopie a původního skenu

Element	Popis	Povinnost	Použití pro
<BasicDigitalObjectInformation>			
<ObjectIdentifier>	údaje o identifikátoru obrazového dokumentu, který je standardem MIX popsán; 0-n	R	MC, PS
<objectIdentifierType>	např. jméno souboru, nebo jiný identifikátor; 0-1	M	MC, PS
<objectIdentifierValue>	hodnota identifikátoru, např. 20110306_001.jp2 nebo urn:nbn:123456; 0-1	M	MC, PS
<fileSize>	velikost souboru 0-1	R	MC + PS
<FormatDesignation>	údaje o formátu obrazového souboru 0-1	M	MC, PS
<formatName>	název formátu, např. lze využít MIME	M	MC, PS

	types[296] (Image/jp2 apod.) 0-1		
<formatVersion>	verze formátu, např. 1.0 0-1	M	MC, PS
<byteOrder>	endianita, možnosti jsou little endian, middle (mix) endian a big endian 0-1	M	MC + PS
<Compression>	údaje o kompresi obrazového souboru (pokud 0-n	M	MC, PS
<compressionScheme>	informace o kompresním schématu, vyjádřeno číslem (např. 34712 je komprese JPEG 2000) nebo slovy (např. JP2 Lossless) 0-1	M	MC, PS
<BasicImageInformation>	základní technické údaje o obrazovém dokumentu 0-1	M	MC, PS
<BasicImageCharacteristics>	0-1	M	MC, PS
<imageWidth>	šířka obrazu v pixelech, např. 3987 0-1	M	MC, PS
<imageHeight>	výška obrazu v pixelech, např. 2345 0-1	M	MC, PS
<PhotometricInterpretation>	photometrická interpretace 0-1	M	MC, PS
<colorSpace>	barevný prostor, např. RGB 0-1	M	MC, PS
<ColorProfile>	údaje o barevném profilu 0-1 povinné pro dokumenty, kde je nutno uchovat přesnou reprezentaci barvy původního dokumentu a používá se ICC profil)	MA	MC + PS
<IccProfile>	ICC profil 0-1	M	MC + PS
<iccProfileName>	jméno profilu, např. sRGB, Adobe RGB aj. 0-1	M	MC + PS
<iccProfileVersion>	verze profilu, např. sRGB IEC61966-2.1 0-1	M	MC + PS
<iccProfileURI>	odkaz na profil, např. www.profily.cz/sRGB_v4_ICC_pref.icc; 0-1	R	MC + PS
<SpecialFormatCharacteristics>	speciální technické údaje o obrazovém	MA	MC

[296] http://www.iana.org/assignments/media-types/index.html

	dokumentu, použití pro formát JPEG 2000 0-1 povinný pro JPEG 2000		
<JPEG2000>	0-1	M	MC
<CodecCompliance>	údaje o kodeku 0-1	M	MC
<codec>	název kodeku, např. Kakadu, LuraWave aj. 0-1	M	MC
<codecVersion>	verze kodeku, např. 3.1 0-1	M	MC
<codestreamProfile >	popis codestream profilu JPEG 2000, např. P0 a P1 (viz ISO/IEC 15444-4); 0-1	M	MC
<complianceClass >	specifikace největší výšky, šířky a počtu komponentů, které dekodér dokáže dekódovat, lze použít hodnoty C0, C1 a C2; 0-1	M	MC
<EncodingOptions >	obsahuje informace o kódování JPEG 2000 0-1	M	MC
<Tiles >	popis pixelové velikosti dlaždic formátu JPEG 2000 0-1	M	MC
<tileWidth>	šířka dlaždice, např. 128 0-1	M	MC
<tileHeight>	výška dlaždice, např. 128 0-1	M	MC
<qualityLayers>	číselná hodnota počtu vrstev, do kterých byl JPEG 2000 rozdělen, např. 12 0-1	M	MC
<resolutionLevels>	popis počtu nižších rozlišení, které lze z obrazu získat, např. 6 0-1	M	MC
<ImageCaptureMetadata>	popis procesu skenování, je důležité vyplnit, protože tyto údaje nelze zjistit z finálního master/archivního souboru 0-1	M	PS
<SourceInformation>	informace o předloze 0-1	R	PS
<sourceType>	Book, Newspaper aj.; nutno používat kontrolovaný slovník 0-1	M	PS
<SourceID>	identifikátor předlohy 0-n	R	PS
<sourceIDType>	typ identifikátoru, např. čČNB, URN:NBN	M	PS

	0-1		
<sourceIDValue>	vlastní hodnota identifikátoru 0-1 povinné	M	PS
<GeneralCaptureInformation>	základní údaje o skenování 0-1	M	PS
<dateTimeCreated>	údaj o datu a čase skenování, např. 2009-01-03T08:25:28; zapsat v ISO 8601 na úroveň vteřin 0-1	M	PS
<imageProducer>	entita provádějící skenování, např. The National Library of the Czech Republic, osoba apod. 0-1	M	PS
<captureDevice>	typ skenovacího zařízení, např. reflection print scanner; doporučené využívání hodnot z kontrolovaného slovníku 0-1	M	PS
<ScannerCapture>	údaje o skeneru 0-1	M	PS
<scannerManufacturer>	výrobce skeneru, např. 4DigitalBooks, Treventus, Zeutschel 0-1	M	PS
<ScannerModel>	údaje o konkrétním typu skeneru 0-1	M	PS
<scannerModelName>	jméno modelové řady skeneru, např. DL 0-1	M	PS
<scannerModelNumber>	číslo/označení modelu, např. 3000 0-1	M	PS
<scannerModelSerialNo>	výrobní číslo skeneru, např. E4R0003649 0-1	M	PS
<MaximumOpticalResolution>	údaje o maximálním optickém rozlišení skeneru 0-1	M	PS
<xOpticalResolution>	optické rozlišení na ose x, např. 300 0-1	M	PS
<yOpticalResolution>	optické rozlišení na ose y, např. 300 0-1	M	PS
<opticalResolutionUnit>	jednotka optického rozlišení, např. inch (in.) 0-1	M	PS
<scannerSensor>	popis typu snímacího senzoru skenovacího zařízení, např. matrix, linear, undefined aj. 0-1	M	PS
<ScanningSystemSoftware>	údaje o softwaru skenovacího zařízení	M	PS

	0-1		
<scanningSoftwareName>	název softwaru, např. Copinet 0-1	M	PS
<scanningSoftwareVersionNo>	číslo verze softwaru, např. 3.7 0-1	M	PS
<DigitalCameraCapture>	údaje o snímacím zařízení (fotoaparát) 0-1 povinné, pokud je používán fotoaparát a není používán skener	MA	PS
<digitalCameraManufacturer>	výrobce fotoaparátu, např. Canon 0-1	M	PS
<DigitalCameraModel>	popis modelu fotoaparátu 0-1	M	PS
<digitalCameraModelName>	název modelové řady, např. EOS 0-1	M	PS
<digitalCameraModelNumber>	označení modelu fotoaparátu, např. 1000D 0-1	M	PS
<digitalCameraModelSerialNo>	výrobní číslo přístroje, např. E12345 0-1	M	PS
<camerarSensor>	typ senzoru fotoaparátu, např. matrix aj. 0-1	M	PS
<CameraCaptureSettings>	údaje o nastavení fotoaparátu použitého ke snímání předloh 0-1	M	PS
<ImageData>	v rámci tohoto kontejnerového elementu budou použity následující sub-elementy: – fNumber – exposureTime – isoSpeedRatings – shutterSpeedValue – apertureValue – brightnessValue – exposureBiasValue – maxApertureValue – subjectDistance – meteringMode – lightSource – flash – focalLength – backLight – exposureIndex – sensingMethod – cfaPattern – autoFocus – PrintAspectRatio	M	PS

	všechny hodnoty budou přebrány v případě použití fotoaparátu z údajů Exif		
<orientation>	popis orientace obrazu tak, jak je uložen vzhledem k jeho řádkům a sloupcům, např. normal*; normal, image flipper; normal, rotated 180°; unknown apod. 0-1	M	PS
<ImageAssessmentMetadata>	informace o digitálním obrazu pro jeho hodnocení a využití z hlediska dlouhodobé ochrany apod. 0-1	M	MC, PS
<SpatialMetrics>	rozměry obrázku, 2 rozměrná projekce objektů tak jak ji „vidí“ snímací zařízení 0-1	M	MC, PS
<samplingFrequencyPlane>	popis základní roviny, např. object plane (pro přímo ze předlohy digitalizované dokumenty), source object plane (pro digitalizaci mikrofilmů), camera/scanner focal plane (indikace sampl. frekvence fyzického senzoru); 0-1	R	MC + PS
<samplingFrequencyUnit>	jednotka měření sampl. frekvence, např. hodnoty 1= žádná pevná jednotka ; 2= inch, 3=centimetr; 0-1	M	MC, PS
<xSamplingFrequency>	údaje o počtu pixelů na jednotku samplovací frekvence pro šířku obrázku 0-1 povinné, pokud hodnota samplingFrequencyUnit je 2 nebo 3	MA	MC, PS
<numerator>	čitatel, číselné vyjádření, např. 300 0-1	M	MC, PS
<denominator>	jmenovatel, číselné vyjádření např. 1 0-1	M	MC, PS
<ySamplingFrequency>	údaje o počtu pixelů na jednotku samplovací frekvence pro výšku obrázku 0-1 povinné, pokud hodnota samplingFrequencyUnit je 2 nebo 3	MA	MC, PS
<numerator>	čitatel, číselné vyjádření, např. 300 0-1	M	MC, PS
<denominator>	jmenovatel, číselné vyjádření např. 1 0-1	M	MC, PS
<ImageColorEncoding>	doplňující údaje o barvě obrazu	M	MC, PS

	0-1		
<BitsPerSample>	počet bitů na kanál 0-1	M	MC, PS
<bitsPerSampleValue>	hodnota počtu bitů, např. 8, 1, 4 nebo 8,8,8 apod. 0-n POZOR – pro každou hodnotu je nutno element opakovat, tj. např. 3x element <bitsPerSampleValue> s hodnotou 8 <mix:BitsPerSample> <mix:bitsPerSampleValue>8</mix:bitsPerSampleValue> <mix:bitsPerSampleValue>8</mix:bitsPerSampleValue> <mix:bitsPerSampleValue>8</mix:bitsPerSampleValue> </mix:BitsPerSample>	M	MC, PS
<bitsPerSampleUnit>	specifikace jednotky, např. integer nebo floating point 0-1	R	MC, PS
<samplesPerPixel>	počet barevných komponentů na pixel, např. 1, 3, 4 0-1	M	MC, PS
<TargetData>	informace o kalibračních tabulkách 0-1 povinné pro obrazy, kde se dělá kontrola oproti kalibrační tabulce	MA	MC
<targetType>	typ kalibrační tabulky; 0= external (kalibrační tabulka se neobjeví na dig. obraze, je to oddělený dig. soubor); 1= internal (tabulka je naskenována spolu s přelohou a objeví se na dig. obraze); 0-n	M	MC
<targetID>	údaje o původu kalibrační tabulky 0-n	M	MC
<targetManufacturer>	výrobce/původce kalibrační tabulky, např. Eastman Kodak nebo NK ČR, oddělení kontroly kvality apod. 0-1	M	MC
<targetName>	název kalibrační tabulky, např. ColorChecker, MicrofilmScanTarget aj. 0-1	M	MC
<targetNo>	číslo nebo verze kalibrační tabulky 0-1	M	MC
<targetMedia>	údaj o tom, na jakém médiu je kalibrační tabulka, např. film, paper aj. 0-1	R	MC
<externalTarget>	údaje o externí kalibrační tabulce; např. link	MA	MC

	na http://skenservis.cz/target-00000001 nebo název a cesta ke konkrétnímu souboru 0-n povinné v případě, že byla použita externí kalibrační tabulka (targetType = 0)		
<performanceData>	odkaz na soubor obsahující charakteristiku výkonu systému vzhledem k nastaveným hodnotám rozlišení atd.; možné hodnoty plnění – link URN nebo URL, nebo název souboru 0-n	R	MC
<ChangeHistory>	dokumentace procesů provedených na obrazovém souboru v jeho životním cyklu 0-1	M	MC
<ImageProcessing>	údaje o zpracování obrazového souboru 0-n	M	MC
<dateTimeProcessed>	2009-01-04T15:12:06; zapsat v ISO 8601 na úroveň vteřin 0-1	M	MC
<sourceData>	odkaz na původní zdrojová data, ze kterých byl vytvořen finální obrazový soubor; může to být např. URL nebo cesta do složky s původním skenem včetně názvu souboru; 0-1	M	MC
<processingAgency>	The National Library of the Czech Republic 0-n	R	MC

8.6 METS část <fileSec>

8.6.1 <fileSec> hlavního záznamu METS

file group

- pro obrazy i texty (ALTO XML/OCR.TXT) budou v hlavním METS záznamu použity elementy <fileGrp>, jeden element <fileGrp> bude existovat pro obrazy archivních kopií, další pro ALTO XML, další pro OCR.TXT soubory a další pro METS záznamy s technickými metadaty (AMD_METS.xml)

1. <fileGrp> pro obrazy archivních kopií, bude mít tyto atributy: ID="MC_IMGGRP" USE="Images"

- každý soubor bude mít vlastní element <file> s následujícími atributy:
  - ID – identifikátor souboru jp2 jak je používán v METS záznamu
  - MIMETYPE – hodnota image/jp2
  - SIZE – velikost souboru jp2

- CHECKSUMTYPE – hodnota MD5
- CHECKSUM – hodnota kontrolního součtu
- SEQ – pořadí souboru
- CREATED – datum vytvoření, ISO8601 na úroveň vteřiny

– subelementem pod <file> je element <Flocat>, který obsahuje link (ideálně v podobě nějakého identifikátoru) na obrazový soubor (xlink:href) a atribut LOCTYPE

2. <fileGrp> pro ALTO XML bude mít následující atributy: ID="ALTOGRP" USE="Layout"

– každý ALTO XML soubor bude mít vlastní element <file> s následujícími atributy:
  - ID – identifikátor souboru ALTO XML jak je používán v METS záznamu
  - MIMETYPE – text/xml
  - SIZE – velikost souboru xml
  - CHECKSUMTYPE – hodnota MD5
  - CHECKSUM - hodnota kontrolního součtu
  - CREATED - datum vytvoření, ISO8601 na úroveň vteřiny

– subelementem pod <file> je element <Flocat>, který obsahuje link (ideálně v podobě nějakého identifikátoru) na xml soubor obsahující ALTO (xlink:href) a atribut LOCTYPE

3. <fileGrp> pro soubory METS s technickými metadaty AMD_METS.xml bude mít následující atributy: ID="TECHMDGRP" USE="Technical Metadata"

– každý METS xml soubor bude mít vlastní element <file> s následujícími atributy:
  - ID - identifikátor souboru AMD_METS.xml jak je používán v METS záznamu
  - MIMETYPE – text/xml
  - SIZE – velikost souboru xml
  - CHECKSUMTYPE – hodnota MD5
  - CHECKSUM - hodnota kontrolního součtu
  - SEQ – pořadí souboru
  - CREATED - datum vytvoření, ISO8601 na úroveň vteřiny

– subelementem pod <file> je element <Flocat>, který obsahuje link (ideálně v podobě nějakého identifikátoru) na xml soubor AMD_METS.xml (xlink:href) a atribut LOCTYPE

4. <fileGrp> pro soubory OCR.TXT bude mít následující atributy: ID="TXTGRP" USE="Text"

– každý OCR.TXT soubor bude mít vlastní element <file> s následujícími atributy:
  - ID - identifikátor souboru OCR.TXT jak je používán v METS záznamu
  - MIMETYPE – text/plain
  - SIZE - velikost souboru
  - CHECKSUMTYPE – hodnota MD5
  - CHECKSUM - hodnota kontrolního součtu
  - CREATED - datum vytvoření, ISO8601 na úroveň vteřiny

– subelementem pod <file> je element <Flocat>, který obsahuje link (ideálně v podobě nějakého identifikátoru) na txt soubor (xlink:href) a atribut LOCTYPE

8.6.2 <fileSec> vedlejšího METS záznamu AMD_METS.xml

- <fileSec> ve vedlejším METS záznamu AMD_METS.xml bude obsahovat jeden element <fileGrp> s vnořenými elementy <file> pro každou reprezentaci stránky, tj. MC, ALTO XML a OCR.TXT
- atributy jednotlivých <file> elementů odpovídají atributům pro jednotlivé typy dokumentů uvedených výše pro <fileSec> hlavního METS záznamu

8.7 METS část <structMap> - Strukturální metadata a ALTO XML

8.7.1 <structMap> hlavního záznamu METS

- strukturální mapy v METS záznamu existují dvojího typu, fyzická a logická; fyzická zaznamenává hierarchické informace o dokumentu, včetně vazeb na fyzické soubory, ze kterých se skládají jednotlivé úrovně dokumentu
- 1 logická strukturální mapa v hlavním METS záznamu popisuje 1 svazek monografie a musí popisovat strukturu až na úroveň vnitřních částí (např. kapitol, nebo článků) apod.
    - součástí svazku monografie mohou být přílohy – pokud se skenují spolu se svazkem, popisuje strukturální mapa METS záznamu svazek včetně přílohy (bere se jako jeden svazek)
- strukturální mapa logická i fyzická včetně linků na ALTO XML bude v hlavním záznamu hlavni_METS.xml
- pro každou stránku seskupuje METS logická strukturální mapa odkazy na textové bloky (nebo ilustrace), které jsou součástí té stránky
    - informace o blocích textu nebo ilustracích na stránce jsou uloženy v 1 ALTO XML souboru, který stránce odpovídá
    - každý blok a každá ilustrace má unikátní identifikátor, který je použit jako odkaz v METS strukturální mapě

Vyjádření fyzické strukturální mapy

- bude mít následující atributy <structMap LABEL="Physical_Structure" TYPE="PHYSICAL">
- fyzická strukturální mapa obsahuje rodičovský , který obsahuje tyto atributy:
    - LABEL- může obsahovat titul svazku monografie
    - TYPE – např. monograph
    - ID – identifikátor div
    - DMDID – identifikátor části popisných metadat náležející ke svazku monografie
- jednotlivé stránky jsou zanořeny do rodičovského elementu jako dceřiné elementy
    - pro soubory stránky bude mít tyto atributy:
        - TYPE – bude se plnit typem stránky (viz typy stránek v DTD periodika http://digit.nkp.cz/Monographs/DTD/2.10/DocumentationMonograph/Monograph.html#element_MonographPage_Link032CD908)
        - ID – identifikátor
        - ORDERLABEL – pořadové číslo stránky, jak je na ní vytištěno
        - ORDER – pořadí stránky ve svazku monografie

- pro soubory stránky vždy obsahují link <ftpr> na soubor obrazu archivní, na ALTO XML, na OCR.TXT a na AMD_METS.xml pomocí elementu <par>
    - link na obrazový soubor archivní kopie má v elementu <area> následující atributy: FILEID, který obsahuje ID souboru archivní kopie
    - link na ALTO XML má v elementu <area> následující atributy: FILEID, který obsahuje ID ALTO XML souboru, dále BEGIN="P1" kde P1 je ID elementu <page> z ALTO XML souboru; a atribut BETYPE="IDREF"
    - link na OCR.TXT soubor má v elementu <area> následující atributy: FILEID, který obsahuje ID souboru OCR.TXT
    - link na AMD_METS.xml soubor má v elementu <area> následující atributy: FILEID, který obsahuje ID souboru AMD_METS.xml

Vyjádření logické strukturální mapy

- bude mít následující atributy <structMap LABEL="Logical_Structure" TYPE="LOGICAL">
- logická struktura na úroveň oddílů nebo např. ilustrací se popisuje pomocí do sebe zanořených elementů
- pokud stránka obsahuje jen obraz a žádný text, pak je popsána jedním elementem s atributem TYPE="PAGE" a link do souboru ALTO XML vede přímo na element <ComposedBlock>
    - lze využít jako kontejner na obrazy a další části stránky, které nejsou součástí článku
    - pro obraz je možno využít atributy a typy podřízených elementů jak je specifikováno v tabulce níže pro PICTURE, který je součástí článku
- stránky obsahující více logických oblastí jsou popsány jedním elementem, který má vnořené elementy pro každou logickou oblast, která odpovídá např. textovému oddílu (např. kapitola, článek) nebo obrazu
    a. pokud se jedná o jednoduchý, celistvý text na jedné straně, tak je popsán jen jedním elementem s atributem TYPE="chapter"
        - v tomto jsou dále jako další elementy zanořeny jednotlivé textové bloky (odstavce, nadpisy, obrazy apod.)
            - u každého bloku je odkaz do ALTO XML souboru na příslušný textový blok <TextBlock> – pomocí tohoto odkazu se v ALTO XML souboru nalezne jak text, tak i informace o jeho umístění na stránce (souřadnice), toto je realizováno pomocí struktury <area> v elementu <ftpr>
            - u bloku tvořeného obrazem je odkaz do ALTO XML na příslušný komponovaný blok <ComposedBlock>; je realizováno pomocí struktury <area> v elementu <ftpr>
            - v případě použití atributu ORDER umožňuje tento princip u oddílů vyjádřit i tzv. pořadí čtení jeho částí, jako jsou např. nadpis, autor, obrázek apod.
    b. výjimečně, pokud textový oddíl není celistvý a je rozdělen na více částí, které se vyskytují na jedné nebo více stránkách, které nemusejí jít za sebou, je možné určit pořadí čtení těchto částí, opět pomocí atributu ORDER

    - pro každou část oddílu existuje vlastní element, podřízený hlavnímu elementu oddílu
    - element každé části má atribut TYPE hodnotu „chapter-part“ a atribut ID musí vyjadřovat, o jakou z částí se jedná, tj. např. ID=“chapter5-1“ odpovídá první části oddílu číslo pět
- **POZOR – u monografie se dělení oddílů běžně nepředpokládá (kapitoly jsou běžně na více stránkách, většinou po sobě jdoucích)**
    - **tj. dělení oddílů není povinné a lze využít pouze struktury odstavců, jak je naznačeno v první části příkladu, tj. pokud kapitola/odstavec pokračuje na další straně, logická mapa uvádí, že poslední odstavec (NORMAL_TEXT) např. na stránce 5 odkazuje na ALTO XML náležející ke stránce 5 a v něm na poslední textový blok; následující odstavec v logické mapě bude odkazovat na ALTO náležející ke straně 6 a v něm na první textový blok (viz příklad)**
- do logické struktury PSP balíčku může být v případě její existence zakomponována i příloha (Supplement), která má vlastní element s atributem TYPE=“SUPPLEMENT“
    - vnořené elementy pro obraz a textové oddíly i jejich použití je shodné se způsobem popisu logické struktury u elementu s atributem TYPE=“VOLUME“

Příklad

Logická mapa obsahující svazek monografie s 5 textovými oddíly (chapter) a 1 přílohou (supplement). První oddíl (kapitola první) je na stránkách 1-2, na straně 1 je titul (název) kapitoly a odstavec normálního textu, na straně 2 je odstavec s normálním textem (může jít také o navazující odstavec – pokračující z předchozí stránky) a obrázek s popiskem i uvedeným autorem). Následují další tři textové oddíly, které nejsou součástí příkladu. Poslední textový oddíl (závěrečná kapitola pátá – doslov) začíná na straně 20 odstavcem s běžným textem (NORMAL_TEXT) a pokračuje na straně 21 dalším odstavcem. Na kapitole páté je ukázána možnost vyjádření návazností jednotlivých rozdělených částí textového oddílu (podobně jako u dělených článků periodik). Tato možnost může být využitelná v případě článků ve sborníku, nebo např. pokud část oddílu (např. kapitoly) pokračuje jinde ve svazku apod.

```
<structMap LABEL="Logical Structure" TYPE="LOGICAL">
    <div LABEL="Honzíkova cesta" TYPE="VOLUME" ID="VOLUME_1" DMDID="XY">
        <div LABEL="Kapitola první - O Honzíkovi" TYPE="CHAPTER" ID="CHAPTER_1" DMDID="XY" ORDER="0">
            <div TYPE="TITLE" ID="CHAPTER_PART_1" ORDER="1">
                <fptr>
                    <area FILEID="ALTO_PAGE_1" BETYPE="IDREF" BEGIN="BLOCK1"/>
                </fptr>
            </div>
            <div TYPE="NORMAL_TEXT" ID="CHAPTER_PART_2" ORDER="2">
                <fptr>
                    <area FILEID="ALTO_PAGE_1" BETYPE="IDREF" BEGIN="BLOCK2"/>
                </fptr>
            </div>
            <div TYPE="NORMAL_TEXT" ID="CHAPTER_PART_3" ORDER="3">
                <fptr>
                    <area FILEID="ALTO_PAGE_2" BETYPE="IDREF" BEGIN="BLOCK1"/>
                </fptr>
            </div>
            <div LABEL="Vesnické nádraží" TYPE="PICTURE" ID="CHAPTER_PART_4" DMDID="XY" ORDER="4">
                <div TYPE="CAPTION" ID="CHAPTER_PART_5">
                    <fptr>
                        <area FILEID="ALTO_PAGE_2" BETYPE="IDREF" BEGIN="BLOCK2"/>
                    </fptr>
                </div>
                <div TYPE="PICT_AUTHOR" ID="CHAPTER_PART_6">
                    <fptr>
                        <area FILEID="ALTO_PAGE_2" BETYPE="IDREF" BEGIN="BLOCK3"/>
                    </fptr>
                </div>
                <div TYPE="IMAGE" ID="CHAPTER_PART_7">
                    <fptr>
                        <area FILEID="ALTO_PAGE_2" BETYPE="IDREF" BEGIN="COMPOSED_BLOCK1"/>
                    </fptr>
                </div>
            </div>
        </div>
        ... následují např. 3 další textové oddíly (kapitoly) až po poslední kapitolu pátou ...
        <div LABEL="Doslov" TYPE="CHAPTER" ID="CHAPTER_5" DMDID="XY" ORDER="1">
            <div TYPE="CHAPTER_PART" ID="CHAPTER_2-1" ORDER="1">
                <div TYPE="NORMAL_TEXT" ID="CHAPTER_PART_1" ORDER="1">
                    <fptr>
                        <area FILEID="ALTO_PAGE_20" BETYPE="IDREF" BEGIN="BLOCK1"/>
                    </fptr>
                </div>
            </div>
            <div TYPE="CHAPTER_PART" ID="CHAPTER_2-2" ORDER="2">
                <div TYPE="NORMAL_TEXT" ID="CHAPTER_PART_1" ORDER="1">
                    <fptr>
                        <area FILEID="ALTO_PAGE_21" BETYPE="IDREF" BEGIN="BLOCK1"/>
                    </fptr>
                </div>
            </div>
        </div>
    </div>
    <div LABEL="Honzíkova cesta" TYPE="SUPPLEMENT" ID="SUPPL_1" DMDID="XY">
        ... popis článků a obrazů stejně jako u TYPE="VOLUME"
    </div>
</structMap>
```

kde jednotlivé části obsahují a popisují…

type	Atributy	Popis	Povinnost
VOLUME nebo SUPPLEMENT	LABEL TYPE ID DMDID	obsahuje údaje o svazku monografie nebo o jeho příloze ---------- LABEL – název (titul) svazku monografie, tedy např. „Honzíkova cesta“ TYPE- hodnota VOLUME nebo SUPPLEMENT ID – identifikátor , např. hodnota „VOLUME_1“ nebo „SUPPL_1“ DMDID – obsahuje identifikátor DMD popisné části MODS svazku/přílohy	M
CHAPTER	LABEL TYPE ID DMDID ORDER	obsahující údaje o jednom textovém oddílu a jeho částech ---------- LABEL – název textového oddílu (např. kapitola, článek ve sborníku apod.) TYPE – hodnota CHAPTER s pořadovým číslem, např. CHAPTER_1 ID – identifikátor elementu DMDID – identifikátor popisných metadat ORDER – pořadí oddílu	M
TYPE=“CHAPTER“ může obsahovat další vnořený různých typů popisující různé části textového oddílu, rozlišujeme tyto části (typy): – TITLE – SUBTITLE – AUTHOR – TRANSLATOR – NORMAL_TEXT – běžný text bez dalšího upřesnění – PICTURE – NOTE – CHAPTER_PART - u oddílů, které jsou rozděleny na více míst na jedné stránce nebo více stránkách (v případě článků ve sborníku např.) - tento pro jednu součást rozděleného článku pak může obsahovat stejné části jako pro oddíl, tj. (TITLE, SUBTITLE, AUTHOR, TRANSLATOR, NORMAL_TEXT, PICTURE)			
TITLE	TYPE ID ORDER	obsahující link na textový blok s nadpisem oddílu (tedy např. kapitoly) ---------- TYPE – hodnota „TITLE“ ID – identifikátor elementu, který popisuje jednu část oddílu (nadpis), např. hodnota „CHAPTER_PART_1“	MA

		ORDER – pořadí části oddílu	
<fptr> <area>	FILEID BEGIN BETYPE	FILEID – ID ALTO XML souboru, např. „ALTO_PAGE_1“ BEGIN – ID textového bloku v ALTO XML souboru BETYPE – hodnota IDREF	
SUBTITLE	TYPE ID ORDER	obsahující link na textový blok s podnadpisem --- TYPE – hodnota „SUBTITLE“ ID – identifikátor elementu, který popisuje jednu část oddílu (podnadpis), např. hodnota „CHAPTER_PART_2“ ORDER – pořadí části oddílu	MA
<fptr> <area>	FILEID BEGIN BETYPE	FILEID – ID ALTO XML souboru, např. „ALTO_PAGE_1“ BEGIN – ID textového bloku v ALTO XML souboru BETYPE – hodnota IDREF	
AUTHOR	TYPE ID ORDER	obsahující link na textový blok se jménem autora --- TYPE – hodnota „AUTHOR“ ID – identifikátor elementu, který popisuje jednu část oddílu (autor), např. hodnota „CHAPTER_PART_3“ ORDER – pořadí části oddílu	MA
<fptr> <area>	FILEID BEGIN BETYPE	FILEID – ID ALTO XML souboru, např. „ALTO_PAGE_1“ BEGIN – ID textového bloku v ALTO XML souboru BETYPE – hodnota IDREF	
TRANSLATOR	TYPE ID ORDER	obsahující link na textový blok se jménem překladatele --- TYPE – hodnota „TRANSLATOR“ ID – identifikátor elementu, který popisuje jednu část oddílu (překladatel), např. hodnota „CHAPTER_PART_3“ ORDER – pořadí části oddílu	MA
<fptr> <area>	FILEID BEGIN BETYPE	FILEID – ID ALTO XML souboru, např. „ALTO_PAGE_1“ BEGIN – ID textového bloku v ALTO XML souboru BETYPE – hodnota IDREF	
NORMAL_TEXT	TYPE ID ORDER	obsahující link na textový blok (nejčastěji odstavec) s běžným textem --- TYPE – hodnota „NORMAL_TEXT“ ID – identifikátor elementu, který popisuje jednu část oddílu (běžný text), např. hodnota „CHAPTER_PART_4“ ORDER – pořadí části oddílu	M
<fptr> <area>	FILEID BEGIN	FILEID – ID ALTO XML souboru, např. „ALTO_PAGE_1“ BEGIN – ID textového bloku v ALTO XML souboru	

	BETYPE	BETYPE – hodnota IDREF	
PICTURE	LABEL TYPE ID DMDID ORDER	pro obraz náležející k textovému oddílu; plní se pokud se obraz vyskytuje --- LABEL – název obrazu pokud existuje TYPE - PICTURE ID – identifikátor elementu, který popisuje jednu část oddílu (běžný text), např. hodnota „CHAPTER_PART_3“ DMDID – link na bibliogr. popis obrazu ORDER – pořadí obrazu	MA
element s typem PICTURE může obsahovat další elementy s typy CAPTION, PICT_AUTHOR, PICT_TITLE a IMAGE; - CAPTION obsahuje text případného popisku k obrazu - PICT_AUTHOR obsahuje text se jménem případného autora obrazu - PICT_TITLE obsahuje text názvu obrazu, pokud nějaký název existuje - IMAGE – obsahuje link do souboru ALTO XML na blok popisující vlastní obraz			
CAPTION	TYPE ID	obsahující link na textový blok s popisem obrazu --- TYPE – hodnota CAPTION ID – identifikátor elementu, např. „CHAPTER_PART_4“	MA
<fptr> <area>	FILEID BEGIN BETYPE	FILEID – ID ALTO XML souboru BEGIN – ID textového bloku v ALTO XML souboru BETYPE – hodnota IDREF	
PICT_AUTHOR	TYPE ID	obsahující link na textový blok s autorem obrazu --- TYPE – hodnota PIT_AUTHOR ID – identifikátor elementu, např. „CHAPTER_PART_5“	MA
<fptr> <area>	FILEID BEGIN BETYPE	FILEID – ID ALTO XML souboru BEGIN – ID textového bloku v ALTO XML souboru BETYPE – hodnota IDREF	
PICT_TITLE	TYPE ID	obsahující link na textový blok s názvem obrazu --- TYPE – hodnota PICT_TITLE ID – identifikátor elementu, např. „CHAPTER_PART_6“	MA
<fptr> <area>	FILEID BEGIN BETYPE	FILEID – ID ALTO XML souboru BEGIN – ID textového bloku v ALTO XML souboru BETYPE – hodnota IDREF	
IMAGE	TYPE ID	obsahující link na komponovaný blok ALTO XML obsahující souřadnice vlastního obrazu ---	MA

		TYPE – hodnota IMAGE ID – identifikátor elementu, např. „CHAPTER_PART_7“	
<fptr> <area>	FILEID BEGIN BETYPE	FILEID – ID ALTO XML souboru BEGIN – ID komponovaného bloku v ALTO XML souboru BETYPE – hodnota IDREF	
NOTE	ID	obsahující link na textový blok s poznámkami k textu --- ID – identifikátor elementu, např. „CHAPTER_PART_9“	MA
CHAPTER_PART	TYPE ID ORDER	obsahující další vnořené odkazující na jednotlivé části konkrétní části rozděleného textového oddílu; možnost použít pro dělený oddíl (typu článek např. ve sborníku) Pozn: pod TYPE=“CHAPTER_PART“ lze vnořit všechny typy jako pod TYPE=“CHAPTER“ --- TYPE – hodnota „CHAPTER_PART“ ID – identifikátor konkrétní části, pro první část děleného oddílu např. „CHAPTER_2-1“, tj. první část oddílu 2 ORDER – pořadí konkrétní části děleného oddílu	MA

Jednotlivé elementy lze kombinovat a vytvářet nové struktury.

8.7.2 <structMap> vedlejšího záznamu METS (AMD_METS.xml)

- bude obsahovat pouze fyzickou strukturální mapu (TYPE=“PHYSICAL“)
- ta bude obsahovat pouze jeden element s atributem TYPE=“MONOGRAPH_PAGE“
- do budou vnořeny odkazy na jednotlivé reprezentace stránky svazku (MC, ALTO XML a OCR.TXT) pomocí elementu <fptr> s atributem FILEID

```
<structMap TYPE="PHYSICAL">
        <div TYPE="MONOGRAPH_PAGE">
            <fptr FILEID="JP2_0001"/>
            <fptr FILEID="ALTOXML_0001"/>
            <fptr FILEID="OCRTXT_0001"/>
        </div>
</structMap>
```

8.8 OCR (ALTO XML a TXT OCR)

- bude použita poslední verze standardu ALTO XML aktuální v době implementace, nebo verze předchozí (prosinec 2010 verze 2 – viz http://www.loc.gov/standards/alto/)
- níže uvedená specifikace **neobsahuje všechny elementy a atributy standardu ALTO XML, obsahuje pouze ty, které jsou pro tuto konkrétní specifikaci relevantní – každý uvedený element má vyjádřenou míru relevance výrazy: povinné, doporučené a nepovinné**
- elementy a atributy, které v této specifikaci nejsou uvedeny, nejsou považovány pro účely specifikace za důležité
- ALTO XML i OCR TXT vzniknou pro všechny obrazové soubory náležející k jedné intelektuální entitě (svazku monografie) včetně prázdných stran, fotografií hřbetu, předsádky apod.
- ALTO XML i OCR TXT budou vznikat na úroveň stránky
- ALTO XML soubor pro zcela prázdné stránky bude obsahovat element /alto/Layout/Page/PrintSpace, ten ovšem nebude obsahovat podelementy[297]
  /alto/Layout/Page/PrintSpace/TextBlock;
  /alto/Layout/Page/PrintSpace/TextBlock/Illustration;
  /alto/Layout/Page/PrintSpace/TextBlock/GraphicalElement ani
  /alto/Layout/Page/PrintSpace/TextBlock/ComposedBlock
- struktura ALTO XML bude generovaná na úrovni rozpoznání slova generovaná OCR
- kvalita rozpoznání znaků bude akceptována do určité hranice, výstupy nebudou ručně opravovány
- struktura ALTO umožní vyhledávání textu a jeho zvýraznění na úrovni slova, pokud bude použit odpovídající prohlížeč
- obrazy reprezentující stránku, které budou použity jako UC, musí odpovídat rozměry, orientací a natočením obrazu, který byl použit pro vytvoření OCR
- OCR TXT bude vznikat z hotových ALTO XML během procesu digitalizace
- ALTO XML se bude vytvářet pouze pro novodobé dokumenty, nebo dokumenty s určitou hranicí kvality OCR
- jméno OCR souboru musí odpovídat jménu obrazového souboru, ke kterému náleží; např. pr_0007.jp2 a al_0007.xml nebo např. 123456_006_alto.xml a 123456_006_archiv.jp2
- kódování ALTO XML i TXT OCR musí být v UTF-8
- souřadnice pozic (HPOS, VPOS, WIDTH, HEIGHT) musí být vyjádřeny v pixelech
- v této specifikaci ALTO XML se počítá s OCR i pro text mimo tzv. textové „zrcadlo“, tj. mimo hlavní text, jako jsou např. čísla stránek, běžící nadpisy ani jiné části vyskytující se na okrajích stránky (top, left, top a bottom margin)
  - elementy <topMargin>, <leftMargin>, <rightMargin>, <bottomMargin> budou obsahovat elementy <TextBlock>, pro které platí stejná pravidla, jako pro element <textBlock> pro hlavní text stránky
  - POZOR: údaje z OCR mimo hlavní text stránky by neměly být vyhledávatelné v aplikaci zpřístupnění, docházelo by ke zmatení uživatele a výsledků (např. při hledání titulu kapitoly by byly zobrazeny výsledky pro každou stránku, která obsahuje běžící nadpis apod.)

[297] V části ALTO XML jsou kvůli větší přehlednosti uvedeny elementy spolu s cestou, která je přesně určuje v rámci schématu (XPath).

- pokud je na konci věty dělící znaménko, ALTO XML i OCR TXT musí obsahovat oba fragmenty slova s dělítkem a současně také kompletní slovo – je vysvětleno dále v tabulce
- ilustrace, reklamy a jiné grafické části stránky nebudou vyjádřeny v tazích /alto/Layout/Page/PrintSpace/Illustration ani Layout/Page/PrintSpace/GraphicalElement, tyto nejsou v popisu/tabulce níže vůbec uvedeny
- ilustrace, reklamy a jiné grafické části stránky budou vyjádřeny v tagu /alto/Layout/Page/PrintSpace/ComposedBlock/ s vyjádřením atributu TYPE, který bude označovat typ bloku (illustration, advertisement aj.)
    - např. ilustrace bude popsána v elementu /alto/Layout/Page/PrintSpace/ComposedBlock/GraphicalElement, kde ComposedBlock TYPE je „Illustration“
    - reklama s textem v rámečku bude popsána v elementu Layout/Page/PrintSpace/ComposedBlock/TextBlock, kde ComposedBlock TYPE je „Advertisement“
    - tabulky, grafy obdobně
- elementy /alto/Layout/Page/PrintSpace/ComposedBlock/Illustration a Layout/Page/PrintSpace/ComposedBlock/ComposedBlock také nebudou využity
- /alto/Layout/Page/PrintSpace/ComposedBlock/TextBlock a /alto/Layout/Page/PrintSpace/ComposedBlock/GraphicalElement nebudou obsahovat elementy <Shape>; tvar těchto bloků je vyjádřen v elementu <Shape> samotného elementu <ComposedBlock>; logicky pak souřadnice tvaru <TextBlock> nebo <GraphicalElement> obsaženého v /alto/Layout/Page/PrintSpace/ComposedBlock jsou většinou shodné, pokud není tvarů nebo bloků v rámci /alto/Layout/Page/PrintSpace/ComposedBlock více
- všechny vyplněné hodnoty jsou příklady plnění, plnění v konkrétní instituci je nutno specifikovat vlastními pravidly a kontrolovanými slovníky
- ALTO XML bude využíváno pro tzv. pořadí čtení, tj. článek vyskytující se na více stránkách nebo na více různých místech jedné stránky bude možné zobrazit celý a ve správném pořadí
    - k tomu je nutno znát jeho strukturu – ta bude vyjádřena v korespondujícím METS záznamu v logické strukturální mapě, která bude obsahovat odkazy na jednotlivé textové bloky článku, pomocí ID textových bloků použitých v ALTO XML

Element	Atribut	Popis	Povinnost
<Description>			
<MeasurementUnit>		měřící jednotka pro souřadnice v ALTO XML; možné hodnoty – dpi, pixel, inch1200 a mm10); inch1200 = 1/1200 inche; doporučené plnění je „mm10“ nebo „pixel“; 0-1	M
<sourceImageInformation>		informace o obrazovém souboru,	M

		ze kterého vzniklo ALTO XML; 0-1	
<fileName>		jméno obrazového souboru, ze kterého bylo ALTO XML vytvářeno; ideálně i s cestou jeho uložení; např. nlaImageSeq-33386-b.tif//produkce/OCR/digibok_XY/XY_011.tiff 0-1	M
<fileIdentifier>		jedinečný identifikátor obrazového souboru; 0-n	R
<OCRProcessing>	ID	popis procesu vzniku OCR; 0-n --- ID OCR procesu, např. <OCRProcessing ID="OCRPROCES_1">; povinné	M
<preProcessingStep>		procesy před vznikem OCR, které provádí SW pro OCR (např. natočení obrazu) 0-n	M
<processingDateTime>		určení času procesu, který předcházel samotnému OCR; např. 2008-03-29T19:42:23 dle ISO 8601 na úroveň vteřin; 0-1	O
<processingAgency>		jméno nebo kód instituce, např. NK CZ, název externí firmy apod.; doporučujeme použít kontrolovaný slovník hodnot; 0-1	R
<processingStepDescription>		popis procesu (např. zarovnání, ořez apod.); 0-n	O
<processingStepSettings>		nastavení kroku popsaného v <processingStepDescription>, např. CCS OCR Processing Filter 0-1	O
<processingSoftware>		popis SW, který upravoval obrázek před vznikem OCR;	M

		0-1	
<softwareCreator>		výrobce softwaru - např. CCS Content Conversion Specialists GmbH, Germany; 0-1	M
<softwareName>		jméno softwaru - např. CCS docWORKS; 0-1	M
<softwareVersion>		verze SW, např. 6.2-1.16; 0-1	M
<ocrProcessingStep>		popis procesu vzniku OCR 1-1 – povinné pole	M
<processingDateTime>		okamžik kdy bylo OCR vytvořeno; nutno zapsat v ISO 8601 na úroveň vteřin; 0-1	M
<processingAgency>		jméno nebo kód instituce, např. NK CZ doporučujeme použít kontrolovaný slovník hodnot; 0-1	M
<processingSoftware>		popis SW, který dělal vlastní OCR; 0-1	M
<softwareCreator>		výrobce softwaru - např. ABBYY, Russia; 0-1	M
<softwareName>		jméno softwaru - např. FineReader; 0-1	M
<softwareVersion>		např. 8.0; 0-1	M
<Styles>		styly definují vlastnosti jednotlivých grafických prvků stránky. styl definovaný v elementu vrchní úrovně je použit jako výchozí pro podřízené elementy; 0-1	M
<TextStyle>	ID FONTSTYLE FONTFAMILY FONTSIZE	definuje font textu; 0-n --- ID pro každý text style použitý v OCR souboru – povinné	M

		FONTSTYLE – např. bold, italics apod.; doporučujeme používat kontrolovaný slovník; doporučené FONTFAMILY – např. arial, calibri apod.; doporučujeme používat kontrolovaný slovník; povinné FONTSIZE – velikost fontu, např. 10, 12 apod.; povinné	
<ParagraphStyle>	ID ALIGN	definuje formátování textových bloků; 0-n --- ID pro každý odstavec + zarovnání; např. PAR_01, PAR_02 apod. povinné ALIGN – zarovnání; povolené hodnoty: Left, Right, Center, Block aj.; povinné	M
<Layout>		layout - rozložení struktur (slov, odstavců apod.) na jedné stránce dokumentu; 1-1 povinný výskyt element není opakovací	M
<Page>	ID ACCURACY POSITION QUALITY PHYSICAL_IMG_NR HEIGHT WIDTH PC	element popisující jednu stránku dokumentu; 1-n --- ID – vygenerovaný identifikátor stránky, např. PAGE1, nebo P1 apod.; povinné ACCURACY – procentuální odhad přesnosti OCR (0-100); doporučené	M

		POSITION – pozice stránky; hodnoty k plnění: Left, Right, Foldout, Single, Cover; nepovinné QUALITY – krátký údaj o kvalitě předlohy stránky; hodnoty k plnění: OK, Missing, Missing in original, Damaged, Retained, Target, As in original; nepovinné PHYSICAL_IMG_NR - fyzické (pořadové) číslo stránky v dokumentu; vyjádřeno číslem, např. 1,2,3 apod.; povinné WIDTH – šířka stránky vyjádřená v pixelech; povinné HEIGHT – výška stránky vyjádřená v pixelech; povinné PC = Confidence level OCR souboru – hodnota mezi 0 (nejistá kvalita) a 1 (dobrá kvalita); nepovinné; pokud nevyplníte ACCURACY – tak je vyplnění doporučené	
<TopMargin>	ID HPOS VPOS WIDTH HEIGHT	horní okraj – prostor mezi vrchní hranou listu a vrchní linkou textu; 0-1 --- ID: unikátní ID pro element TopMargin, např. P1_TM0001 (page 1, topMargin0001); povinné HPOS: horizontální pozice; povinné	M

		VPOS: vertikální pozice; povinné WIDTH – šířka vrchního okraje; povinné HEIGHT – výška vrchního okraje; povinné	
<TextBlock>	stejné plnění a pravidla jako pro element <TextBlock> vnořený do elementu <PrintSpace>		MA
<LeftMargin>	ID HPOS VPOS WIDTH HEIGHT	levý okraj – prostor mezi levým okrajem stránky a textem; 0-1 ---------- ID: unikátní ID pro element LeftMargin, např. P1_LM0001 (page 1, leftMargin0001); povinné HPOS: horizontální pozice; povinné VPOS: vertikální pozice; povinné WIDTH – šířka levého okraje; povinné HEIGHT – výška levého okraje; povinné	M
<TextBlock>	stejné plnění a pravidla jako pro element <TextBlock> vnořený do elementu <PrintSpace>		MA
<RightMargin>	ID HPOS VPOS WIDTH HEIGHT	pravý okraj – prostor mezi pravým okrajem stránky a textem; 0-1 ---------- ID: unikátní ID pro element RightMargin, např. P1_RM0001 (page 1, rightMargin0001); povinné HPOS: horizontální pozice; povinné	M

		VPOS: vertikální pozice; povinné WIDTH – šířka pravého okraje; povinné HEIGHT – výška pravého okraje; povinné	
<TextBlock>	stejné plnění a pravidla jako pro element <TextBlock> vnořený do elementu <PrintSpace>		MA
<BottomMargin>	ID HPOS VPOS WIDTH HEIGHT	pravý okraj – prostor mezi spodním okrajem stránky a textem; 0-1 --- ID: unikátní ID pro element BottomMargin, např. P1_BM0001 (page 1, bottomMargin0001); povinné HPOS: horizontální pozice; povinné VPOS: vertikální pozice; povinné WIDTH – šířka spodního okraje; povinné HEIGHT – výška spodního okraje; povinné	M
<TextBlock>	stejné plnění a pravidla jako pro element <TextBlock> vnořený do elementu <PrintSpace>		MA
<PrintSpace>	ID HPOS VPOS WIDTH HEIGHT	popis tvaru pokrývajícího textové pole stránky; 0-1 --- ID: unikátní ID pro element <printSpace>, např. P1_PS0001 (page 1, printSpace0001); - povinné HPOS: horizontální pozice; povinné	M

		VPOS: vertikální pozice; povinné WIDTH – šířka textového pole; povinné HEIGHT – výška textového pole; povinné	
<TextBlock>	ID STYLEREFS HPOS VPOS WIDTH HEIGHT	popisy textových bloků na konkrétní stránce; 0-n pokud je stránka prázdná, TextBlock není potřeba uvádět; pokud je na stránce text tak ano ---------- ID obsahuje identifikátor textového bloku na stránce, např. "BLOCK1" nebo P1_TB0002 (stránka 1, textový blok 2); povinné STYLEREFS: reference na ID definice formátování textových bloků <ParagraphStyle>; povinné HPOS: horizontální pozice bloku; povinné VPOS: vertikální pozice bloku; povinné WIDTH – šířka textového bloku; povinné HEIGHT – výška textového bloku; povinné	MA
<Shape>		tvar textového bloku; 0-1 – pro jeden výskyt <TextBlock> jeden nebo žádný výskyt <Shape>; plnit v případě, že je tvar textového bloku nestandardní (víceúhelník)	RA

<Polygon>	POINTS	popis (souřadnice) tvaru víceúhelníku; 0-1 --- POINTS – vyjádření jednotlivých bodů víceúhelníku; povinné	M
<TextLine>	ID STYLEREFS HPOS VPOS WIDTH HEIGHT	popis jedné řádky textu v rámci textového bloku; 1-n nutný alespoň jeden výskyt v rámci textového bloku --- ID obsahuje identifikátor řádky textu v textovém bloku, např. "P1_TL0002 (stránka 1, řádka 2); povinné STYLEREFS: reference na ID definice formátování textových bloků <ParagraphStyle>; nepovinné HPOS: horizontální pozice řádky; povinné VPOS: vertikální pozice řádky; povinné WIDTH – šířka řádky; povinné HEIGHT – výška řádky; povinné	M
<String>	ID CONTENT HEIGHT WIDTH HPOS VPOS CC WC V případě dělení slov také:	řetězec znaků – vlastní obsah OCR; znaky tvoří jednotlivá slova a více tagů <String> větu <TextLine>; 1-n v rámci <TextLine> --- ID obsahuje unikátní sekvenční číslo řetězce na stránce, např. "P3_ST0001" (strana 3, řetězec 1); povinné	M

	SUBS_TYPE SUBS-CONTENT	CONTENT – ukládá vlastní řetězec znaků (slovo); povinné HPOS: horizontální pozice řetězce; povinné VPOS: vertikální pozice řetězce; povinné WIDTH – šířka řetězce; povinné HEIGHT – výška řetězce; povinné CC – úroveň důvěry v přesnost OCR rozpoznání každého znaku v řetězci; jde o seznam čísel, každé z nich mezi hodnotami 0 (jistá) a 9 (nejistá) pro každý znak; např. CC=“0001“ pro CONTENT=“TEXT“; povinné WC – úroveň důvěry v přesnost OCR výstupu celého řetězce - slova (word confidence); hodnota mezi 0 (nejistá) a 1 (jistá); např. WC=“0,99“; povinné SUBS_CONTENT – obsah chybějící části řetězce v případě, že je slovo na konci řádku rozdělené i do druhého řádku; obsahuje celý řetězec - aby byl vyhledatelný i v případě, že slovo se na stránce vyskytuje, ale je rozděleno; povinné	

		SUBS_TYPE – označení typu substituce; možné hodnoty: HypPart1; HypPart2; Abbreviation; povinné - při výskytu SUBS_CONTENT *HypPart1* se vyskytuje při rozdělení slova u jeho první OCR části (u první části tagu <CONTENT> ve větě (stringu) první; *HypPart2* se vyskytuje u následujícího tagu <CONTENT> v následující větě (stringu), který obsahuje druhou část rozděleného slova/řetězce; *Abbreviation* – typ substituce používaný při rozepisování zkratek v textu na jejich plný text; při dělení slov v textu HypPart1 a HypPart2 povinné, abbreviation nepovinné	
<ALTERNATIVE>		alternativní hodnota OCR řetězce pro jednotlivá slova; 0-n lze použít v případě nejistoty rozpoznání řetězce;	O
<HYP>	CONTENT WIDTH HPOS VPOS	zápis znaku rozdělovníku slov 0-1 pro jeden výskyt <TextLine>; vždy pro poslední <String>; může se vyskytnout pouze na konci řádku (1x) --- CONTENT – obsahuje řetězec znaků, které jsou v textu použity na rozdělení slova, nejčastěji „-„; povinné WIDTH – šířka dělícího znaku; doporučené HPOS: horizontální pozice	MA

		dělícího znaku; doporučené VPOS: vertikální pozice dělícího znaku; doporučené	
<SP>	ID WIDTH HPOS VPOS	prázdný prostor mezi řádky; 0-n v rámci jednoho <TextLine>; vždy mezi řádky, tj. mezi tagy <String>; --- ID: unikátní ID pro prázdný prostor mezi řádky, např. P1_SP0001 (stránka 1, prázdný prostor 0001); povinné HPOS: horizontální pozice; povinné VPOS: vertikální pozice; povinné WIDTH – šířka prázdného prostoru; povinné	M
<ComposedBlock>	ID TYPE HPOS VPOS WIDTH HEIGHT STYLEREFS	blok sestávající z jiných bloků; může obsahovat PrintSpace/ComposedBlock/TextBlock, PrintSpace/ComposedBlock/Illustration, PrintSpace/ComposedBlock/GraphicalElement, /PrintSpace/ComposedBlock/ComposedBlock, tj. stejné elementy (bloky), které obsahuje samotný element /alto/Layout/Page/PrintSpace; 0-n povinné pro vyjádření bloků textu (např. orámovaný text, reklamy), pro vyjádření ilustrací, tabulek a	MA

		grafik --- ID: unikátní ID komponovaný blok, např. P6_CB0001 (stránka 6, komponovaný blok 0001); povinné TYPE – označení typu komponovaného bloku; nutné používat kontrolovaný slovník (illustration, Advertisement, apod.); povinné HPOS: horizontální pozice bloku; povinné VPOS: vertikální pozice bloku; povinné WIDTH – šířka komponovaného bloku; povinné HEIGHT – výška komponovaného bloku; povinné	
<Shape>		tvar komponovaného bloku; 0-1 – pro jeden výskyt /alto/Layout/Page/PrintSpace/ComposedBlock jeden nebo žádný výskyt /alto/Layout/Page/PrintSpace/ComposedBlock/Shape; doporučeno – v případě, že je tvar komponovaného bloku nestandardní (víceúhelník)	RA
<Polygon>	POINTS	popis tvaru víceúhelníku; 0-1 --- POINTS – vyjádření jednotlivých bodů víceúhelníku povinné	M

<TextBlock>	ID STYLEREFS HPOS VPOS WIDTH HEIGHT	v případě, že komponovaný blok (např. orámovaný tvar) obsahuje text; platí stejná pravidla jako pro normální element /alto/Layout/Page/PrintSpace/TextBlock; 0-n (pro jeden výskyt <ComposedBlock> 0 nebo více elementů /alto/Layout/Page/PrintSpace/ComposedBlock/TextBlock>; plnit pokud je v komponovaném bloku text ---------- ID obsahuje identifikátor textového bloku v komponovaném bloku, např. P1_CB0002_SUB (stránka 1, textový blok 2, SUB značí komponovaný blok); povinné STYLEREFS: reference na ID definice formátování textových bloků /alto/Styles/ParagraphStyle; povinné HPOS: horizontální pozice bloku; povinné VPOS: vertikální pozice bloku; povinné WIDTH – šířka textového bloku; povinné HEIGHT – výška textového bloku; povinné	MA
<TextLine>	/alto/Layout/Page/PrintSpace/ComposedBlock/TextBlock/TextLine a ostatní elementy v rámci /alto/Layout/Page/PrintSpace/ComposedBlock/TextBlock mají stejná pravidla a výskyty jako jako ve vrchním elementu /alto/Layout/Page/PrintSpace/TextBlock		

<GraphicalElement>	ID HPOS VPOS WIDTH HEIGHT	popis grafického tvaru; v případě využití v rámci /alto/Layout/Page/PrintSpace/ComposedBlock označuje rozměry tvaru v rámci něhož je tabulka, ilustrace, reklama apod.; 0-1 - pro jeden výskyt /alto/Layout/Page/PrintSpace/ComposedBlock 0 nebo max. 1 výskyt <GraphicalElement>; plní se, pokud je na stránce a tedy v komponovaném bloku ilustrace, tabulka apod.; ---------- ID – identifikátor grafického tvaru; povinné HEIGHT – výška grafického tvaru; povinné WIDTH – šířka grafického tvaru; povinné HPOS – horizontální pozice grafického tvaru; povinné VPOS – vertikální pozice grafického tvaru; povinné	MA