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

Teze disertační práce

Vedoucí práce – Stanislav Kalkus, Ph.D.

2012

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 (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 vytváří také metadata v projektu NK ČR ANL+ (analytické zpracování periodik).

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.

1. Předmluva

Dnešní společnost již není společností „papírovou“, nýbrž elektronickou. Velká část vědění vzniká pouze v elektronické podobě a část dokumentů, která byla tradičně uložena na papírových nosičích, je nyní přístupná také v digitální podobě. Tato skutečnost se zdá býti výhodnou, hlavně uživatelsky, ovšem je třeba si uvědomit, že ve své podstatě představuje velkou hrozbu. Hrozbu nenávratné ztráty digitálních informací, vědění, zkušeností a poznatků. V posledních letech si pracovníci tzv. paměťových institucí (knihovny, archivy, muzea) začali uvědomovat, že nebezpečí je velmi reálné a že ztráty digitálních informací a relevantních dokumentů jsou aktuálním problémem, kterému musí čelit.
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 fundamentální změny těchto procesů. Předkládaná disertační práce popisuje mj. tento společenský a technologický jev a to, jak se s ním vyrovnávají knihovny a archivy ve světě i u nás. V posledních letech 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. U vyspělých systémů na ukládání dat logická dlouhodobá ochrana stojí a padá s metadaty, která jsou v systému uložena. Jde nejen o metadata popisná, ale především o strukturální, administrativní, ochranná a technická. Právě o těchto metadatech je předkládaná disertační práce.
Práce je zacílena na digitální dokumenty vzniklé digitalizací. Jsou popsány základní problémy dlouhodobé ochrany digitálních dat (digital preservation) i nabízející se řešení. Jedním z nejdůležitějších způsobů jak zajistit dlouhověkost, použitelnost a pochopitelnost digitálních objektů v budoucnu, je opatřit je odpovídajícími údaji o nich samotných, tj. právě metadaty. Jsou popsány jednotlivé typy metadat, historie jejich vzniku a využívání ve světě. Podrobněji je pojednáno z různých hledisek o využití metadat v projektech NK ČR.
Hlavní praktickou částí disertační práce je návrh nového aplikačního profilu metadat pro digitalizaci periodik a monografií v projektu *Národní digitální knihovna* (NDK). Bylo nutné popsat vývoj a stav využití metadat v obou dosavadních projektech NK ČR, které se zabývají digitalizací (*Kramerius* a *Manuscriptorium*). Popis vývoje je východiskem návrhu nových profilů metadat. 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. Disertační práce obsahuje také přiblížení procesu rozhodování a implementace nových typů metadat v procesu digitalizace projektu NDK. Návrh nového profilu obsahuje nejen popisná, ale především administrativní, technická, strukturální a ochranná metadata. Návrh aplikačního profilu metadat pro digitalizaci v projektu NDK je specifikován a vytvořen tak, aby byly digitální dokumenty vycházející z tohoto projektu připraveny na dlouhodobou ochranu v rámci long-term preservation (LTP) systému.

Disertační práce se také věnuje vývoji způsobu uložení dat a digitálním repozitářům (digital repositories). Konkrétně možnostmi jejich implementace a správou dokumentů v dlouhodobém horizontu tak, aby digitální dokumenty v nich uložené byly stále uživatelům dostupné, použitelné a také pochopitelné. Zjednodušeně řečeno, aby např. dokumenty vzniklé v určitém SW v roce 1995 byly k dispozici v použitelné podobě např. i v roce 2055 bez větších komplikací, tak jako je tomu dnes v případě knih. V této souvislosti je rozebrána problematika vnitřních auditů a externí certifikace repozitářů, které by měly prokázat opravdovou spolehlivost institucí a jejich repozitářů.

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.

Následující text obsahuje původní teze disertační práce, jak byly předloženy v roce 2008. Text pod tezemi reflektuje obsah výsledné předložené práce a její závěry.

2. Digitální dokumenty a dlouhodobá ochrana digitálních dat

Již bylo zmíněno v úvodu, že nahrazování analogových dokumentů digitálními přináší vedle výhod také rizika. Konkrétně se tato rizika pojí se zastaráváním HW a SW, které může způsobit, že digitální dokumenty z určité doby budou v budoucnu naprosto nesrozumitelné a nepoužitelné. Může se stát, že budoucí uživatel nebude mít možnost si digitální objekt zobrazit na konkrétním HW. Toto se ostatně již děje[1]. To samé platí i pro SW aplikace. Bez opatření tzv. logické dlouhodobé ochrany nemusí být snadné ani rozpoznat, zda se v případě konkrétního digitálního objektu jedná o text, audio nebo např. o obrazový dokument. Problém zastarávání HW a hlavně SW se stává aktuálním ve všech oblastech lidské činnosti, kde je potřeba uchovávat digitální objekty v dlouhodobém horizontu. Oblast ochrany digitálních objektů je v současnosti velmi úzce spojená s problematikou velkých digitálních repozitářů.

Teze 1.

V posledních letech je stále více a více dokumentů vytvářeno v digitální podobě. Paměťové instituce jsou postaveny před problém uložení a dlouhodobé ochrany digitálních dokumentů (vzniklých v procesu digitalizace, případně digital-born).

- Archivované digitální objekty musí být použitelné uživateli, kteří jsou vzdáleni v čase, prostoru a nemají podporu producenta konkrétní informace/objektu.
  - Producent informace již neexistuje, nemůže odpovědět na žádné otázky. Jediné co máme, je digitální objekt a jeho metadata.
  - SW, na kterém byla informace (digitální objekt) vytvořena, již nemusí být podporován. Informace zaznamenané tímto SW mohou být zcela nedostupné (neznámé kódování, nedostupná dokumentace).
- Uživatelská komunita se mění během doby. Nová komunita nebude znát pozadí konkrétních informací (účel vzniku, SW, HW) a může používat naprosto odlišné pracovní prostředí.
- Digitální dokumenty (objekty) jsou zranitelné skrz svůj formát a skrz technologie potřebné pro jejich zobrazení.
- Právě formáty dat, jejich různorodost, rychlý vývoj a následná nekompatibilita s novými SW (zastarávání SW) je jedním z hlavních problémů oblasti dlouhodobé ochrany digitálních dat.
- Udržení zpřístupnitelnosti digitálních objektů je závislé na SW a HW a na odpovídajících metadatech, které dokumenty doprovázejí.
- Možnosti řešení logické i základní dlouhodobé ochrany digitálních dat existují. Mezi nejužívanější patří migrace a emulace.

Teze 2.

Zájem a pozornost komunity se přesouvá od vytváření digitálních objektů (digitalizace) k dlouhodobému uchování digitálních objektů ve specializovaných systémech.

- Mnoho organizací, které digitálních informací již delší dobu využívají nebo digitalizují své sbírky, se nyní dostává do situace, kdy musí chtě nechtě přistoupit k archivním opatřením pro digitální objekty, které ukládají. To platí i pro NK ČR, kde se problematika v teoretické rovině řeší až od roku 2008. Snahy vyústily v plán projektu NDK.

[1] Vzpomeňme např. náhradu děrných štítků za magnetické pásky, posléze za pevné paměti, optické disky apod.

- Všudypřítomnost otázek spojených s logickou dlouhodobou ochranou digitálních dat vytváří obecné základy pro mezioborový dialog a spolupráci pro lepší odpovědi na všechny výzvy.
- Evropská unie snahy o řešení problémů dlouhodobé ochrany digitálních dat podporuje ve svých projektech od Rámcového programu 5 až dosud. Lze říci, že na úkor digitalizace, která již nemá takovou podporu, jako před několika lety. Výstupy projektů EU v posledních letech nabízejí reálné nástroje a metody, které jsou v práci popsány.
- Projekty digitalizace musí dnes obsahovat i část věnovanou logické dlouhodobé ochraně v projektu vzniklých dat, udržitelnosti v budoucnu a ochranným metadatům.
- Důvodem snah o logickou ochranu digitálních dat je mnohdy zákonná povinnost (knihovny, archivy), nebo např. ochrana financí vložených do digitalizace.

Teze 3.

Mnoho institucí si stále neuvědomuje archivní funkci v rámci svých mandátů, nebo na ochranu digitálních objektů nemají odpovídající znalosti ani finance; hrozí tak nenávratná ztráta jejich digitálních dokumentů.

- Naléhavost a potřeba k podniknutí prvních kroků ochranných opatření k zajištění dlouhodobé životnosti digitálního materiálu zasáhla všechny instituce – instituce paměťové, podniky, vládní úřady atd.
- Knihovnická komunita v ČR si stále ne zcela uvědomuje, že se i přes poměrnou novost a aktuálnost digitálních objektů, se jedná o budoucí kulturní dědictví, které má nárok na odpovídající ochranu.
- Ochránit a zachovat digitální data v dlouhodobém horizontu je daleko těžší a finančně náročnější než např. v případě papíru. Dlouhodobá archivace a následně logická ochrana digitálního objektu představuje náročný úkol doprovázený celou řadou problémů [GIARETTA, [2009]].
- Uložení klasických dokumentů bylo statické, a třebaže si žádalo nějaké výlohy, představovaly jen nepatrný zlomek ekonomické náročnosti na dynamickou archivaci, tj. údržbu a migraci elektronických informačních pramenů. Nejedná se pouze o investice do technologií, daleko více finančně náročné je udržovat vhodné prostředí pro repozitář, personalistika, mzdové prostředky, zálohování, apod. 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. 87].
- Je popsána existující projektová podpora problematiky logické dlouhodobé ochrany digitálních dat a účast NK ČR v těchto projektech.

3. Obecné mechanismy zajištění logické dlouhodobé ochrany digitálních objektů

Mnoho institucí v ČR stále neakcentuje rozdíl dlouhodobé ochrany digitálních dat ve smyslu pouhých záloh (ochrana bitstreamu) a tzv. logické dlouhodobé ochrany, která spočívá v migracích formátů, případně emulacích, vytváření metadat, doplňování metadat, hlídání rizik spojených s formáty – to vše tak, aby bylo možné na rizika reagovat a digitální objekt zachovat vyhledatelný, zobrazitelný, pochopitelný i v budoucnu. Logická ochrana reaguje na technologické změny a může případně digitální objekt měnit (formát). Při tom všem je nutno zachovat autenticitu objektu. Ochrana bitstreamu může objekt přesouvat a zálohovat na nová média, ovšem objekt jako takový zůstává stále stejný, což bude působit problémy s jeho využitím v budoucnu. Ochrana bitstreamu je prvním

předpokladem pro logickou dlouhodobou ochranu digitálních dat.
Obecnými mechanismy pro logickou dlouhodobou ochranu je tvorba ochranných, administrativních, strukturálních a technických metadat a jejich využívání při sledování životního cyklu digitálního objektu, plánování ochrany a vlastních ochranných akcích. Těmito ochrannými akcemi jsou ve zjednodušeném pohledu migrace formátů a/nebo emulace HW a SW prostředí.
Digitální objekty jsou v současnosti uchovávány v digitálních repozitářích a specializovaných systémech pro logickou dlouhodobou ochranu (LTP systém). Máme-li hledat odpověď na otázku, proč jsou vybudování a provoz digitálního repozitáře finančně i personálně natolik náročné, je třeba pochopit jeho základní funkce. V úplnosti je mapuje (a do detailů rozvíjí) referenční model OAIS (Open Archival Information system). Úsilí vyvinout archivní standardy pro dlouhodobé ukládání dat v digitální formě vzniklo na základě potřeby vývoje datového standardu pro podporu ochrany dat vzniklých ve vesmírném výzkumu. Výzkum byl zadán organizaci Consultative Committee for Space Data Systems (CCSDS), což je orgán pro mezinárodní spolupráci vesmírných agentur. CCSDS navrhlo 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 a nakonec měl posloužit jako rámec pro další aktivitu. Referenční model OAIS byl zveřejněn v roce 1999 a okamžitě se rozšířil i mimo původní oblast. Je využíván pro architekturu repozitářů a široce implementován v paměťových institucích.

Teze 4.

Úspěšný a funkční digitální archivní systém (repozitář) je nutné postavit na základě referenčního rámce OAIS.

- Referenční model OAIS je koncepční rámec pro archivní systém věnovaný problému ochrany a správy přístupu k digitálním informacím v dlouhodobém měřítku. Popisuje prostředí, ve kterém digitální archiv (repozitář) sídlí, jeho funkční komponenty a informační infrastrukturu podporující všechny procesy archivu [OCLC/RLG WORKING GROUP ON PRESERVATION METADATA. 2002, s. 5].
- Referenční model OAIS představuje obecný (high level) popis typů informací vytvářených a spravovaných funkčními komponenty celého archivního systému. Nezabývá se konkrétními typy digitálních objektů, které jsou spravovány v archivu, ani specifikací technologie nasazené v archivu/repozitáři, aby se dosáhlo cíle, tj. ochrany a údržby přístupu k digitálním objektům v dlouhodobém horizontu.
- Referenční model OAIS se ukázal jako velmi životaschopný a je dnes široce implementován a využíván. Je ideální svou obecností a tím tedy i možností implementace. Celý projekt NDK v NK ČR stojí na implementaci referenčního rámce OAIS. Jeho použití je nutné pro architekturu LTP systému a pro návrh aplikačního profilu metadat pro proces digitalizace. Kompatibilita s OAIS je klíčová pro interoperabilitu dat i metadat, možnosti využití dostupných nástrojů a konceptů. Odmítnutí referenčního modelu by znamenalo izolaci ve snaze o dlouhodobou ochranu digitálních dat i v oblasti zpřístupnění.
- Struktura OAIS je podporována většinou využívaných metadatových standardů i archivních SW/systémů, což mnoho věcí ulehčuje, hlavně při vzájemné kooperaci.
- Referenční model OAIS je výchozím bodem vzniku a konceptu současných standardů technických, ochranných a administrativních metadat. Z tohoto pohledu je blíže vysvětlen.

Teze 5.

Implementace konceptu tří „informačních balíčků“ jako součásti archivu odpovídajícího OAIS je z hlediska dalšího využití SW, implementace metadat a interoperability digitálního repozitáře klíčová.

- Repozitář odpovídající OAIS pracuje s několika typy informačních balíčků. Ty mohou obsahovat informace o obsahu a informace podpůrné (metadata). Tyto informační balíčky pak tvoří jeden z balíčků, které mají konkrétní úkoly v rámci repozitáře. Jsou to SIP, AIP a DIP (Submission, Archival a Dissemination Information Package).
- Informační balíčky jsou v OAIS archivu přesouvány mezi jednotlivými moduly jak je OAIS specifikuje. Tyto moduly jsou popsány s použitím příkladů ze specifikace LTP systému založeného na OAIS z projektu NDK.
- Pro uvedené tři typy balíčků se používají různé úrovně metadatového popisu, nejpodrobnější je archivní balíček AIP, který obsahuje všechny typy doprovodných informací, jak je definuje referenční model.
- Návrh profilu metadat pro projekt NDK je návrhem balíčku SIP, který obsahuje metadata pro další použití v balíčku AIP[2].

Teze 6.

Nejčastěji používané ochranné akce pro logickou digitální ochranu digitálních objektů jsou formátová migrace a emulace.

- Migrace je rozšířenější a je podporována mnoha nástroji a komplexními systémy.
- Emulace jako ochranná akce se rozvíjí zejména v posledních letech. Počítá se s ní na logickou ochranu velmi komplexních dokumentů, např. webových stránek nebo databází.
- Popsány jsou výhody a nevýhody obou přístupů, včetně dalších možností řešení dlouhodobé ochrany digitálních dat.
- Disertační práce popisuje nejčastěji užívané nástroje na tvorbu metadat, validaci formátů dat, identifikaci formátů dat, uložení digitálních objektů, migrace, emulace apod.

4. Metadata pro dlouhodobou ochranu digitálních objektů

Problematika metadat využívaných v digitálních repozitářích je již několik let velmi aktuální. Na metadatech závisí, zda budeme schopni naše digitální objekty využívat v budoucnu, zda budou přístupné, bude možné je přečíst, zobrazit, budeme si jisti, že jsou data autentická apod. Tuto skutečnost si uvědomuje většina světových paměťových i vědeckovýzkumných institucí, které mají zájem na dlouhodobém uchování svých dat a sbírek pro budoucí uživatele. Vědí, že je to z možných cest ta jednodušší, i když také velmi náročná.

Logická dlouhodobá ochrana digitálních dat je na metadatech ve své dnešní podobě závislá. Vyžaduje nejen popisné, ale hlavně technické, administrativní a ochranné informace (metadata) ke konkrétnímu digitálnímu objektu. Na základě těchto informací poté relevantní systémy nebo i osoby jsou schopné zhodnotit rizika, která konkrétnímu digitálnímu objektu hrozí a rozhodnout se o případných akcích ochrany (preservation action). Bez metadat by to možné nebylo. Disertační práce popisuje metadata obecně a věnuje se detailně jednotlivě výše uvedeným typům metadat. Posuzuje je z pohledu významu pro logickou dlouhodobou ochranu. Kapitola o metadatech obsahuje i

[2] Návrh aplikačního profilu projektu NDK ovšem (záměrně) nepokrývá metadata AIP balíčku.

rozsáhlou část o vývoji relevantních standardů metadat.
Zatímco u klasických dokumentů je relativně snadné zajistit jejich autenticitu a jejich ohrožení lze zjistit pouhým okem při prohlídce skladiště, u digitálních dokumentů je obojí podstatně složitější, stejně jako jejich zajištění proti neoprávněnému užití. Správa vlastních digitálních objektů i souvisejících metadat je složitý a permanentní proces.
Digitální repozitář přijímá digitální objekty od producentů, zpravidla včetně dohodnutých metadat, v podobě tzv. SIP balíčku. Tento SIP balíček musí obsahovat již výše jmenované typy metadat tak, aby podporovala následné procesy ochrany. Metadata z balíčku SIP jsou prvotní informací o digitálním objektu, o jeho vzniku, procesech a technických vlastnostech. Ty jsou vstupem pro tzv. systém na dlouhodobou ochranu (LTP systém), který s nimi dále dokáže pracovat, dále je obohacovat a aktualizovat. Je schopen tato metadata využít na automatizované analýzy rizik a na procesy plánování dlouhodobé ochrany. S tímto záměrem vznikl i návrh nových profilů metadat pro digitalizaci v projektu NDK.

Teze 7.

Pro logickou dlouhodobou ochranu digitálních objektů je nezbytné opatřit je tzv. ochrannými metadaty.

- Termín „ochranná metadata“ zahrnuje několik kategorií obyčejně užívaných k rozlišení typů metadat. Jsou to metadata: administrativní (včetně práv a povolení); technická a strukturální. Někdy se termín používá samostatně, vedle výše uvedených.
- Ochranná metadata jsou informační strukturou, která podporuje procesy spojené s logickou ochranou digitálních dokumentů. Konkrétněji, jsou informací nutnou k zachování životnosti, „zobrazitelnosti“ a srozumitelnosti digitálních objektů v dlouhodobém horizontu.
  - Životností je myšleno, že bitstream archivovaných digitálních objektů je neporušený a čitelný z média, na kterém je uložen.
  - Zobrazitelnost odkazuje na schopnost převedení bitstreamu do formy, kterou může uživatel nebo SW, případně aplikace bez problémů číst.
  - Srozumitelnost označuje poskytování dostatečného množství informací, tj. že zobrazovaný obsah je pochopitelný cílovému uživateli.
- Ochranná metadata jsou nezbytná pro emulaci nebo migraci dokumentů.
- Ochranná metadata jsou nutným vstupem pro LTP systém. I v případě, že výsledek digitalizace bude pouze uložen v běžné digitálním repozitáři a ne specializované LTP systému, je tento typ metadat klíčový pro pozdější pochopení vlastností digitálního objektu a možnosti kontroly jeho integrity, autenticity.

Teze 8.

Většině digitálních objektů ukládaných dnes v digitálních repozitářích (jakéhokoliv typu) chybějí právě ochranná metadata, která jsou pro dlouhodobou ochranu digitálních dokumentů tak důležitá.

- Je popsán vývoj používání metadat pro digitalizaci novodobých dokumentů v projektu *Kramerius* (VISK7). Dodnes používaná DTD monografie a periodika byla vytvořena v roce 2003 mj. na základě projektu DIEPER (Univerzitní knihovna v Göttingen) a předchozích standardů metadat DOBM. Bohužel obsahují hlavně popisná metadata, stejně jako metadatový popis digitálních objektů v jiných projektech v ČR.
- Podobně je z tohoto pohledu rozebrán a komentován vývoj metadat používaných v projektu

digitalizace historických dokumentů *Manuscriptorium*.

- Pro potřeby dlouhodobé ochrany digitálních dat je do procesů digitalizace v NK ČR doplnit další údaje o digitálních objektech, konkrétně administrativní a technická metadata. Pro novodobé dokumenty je provedeno v projektu NDK v rámci nového profilu metadat.

Teze 9.

Digitální objekty by už na vstupu do digitálního repozitáře měly být doprovázeny technickými, ochrannými a administrativními metadaty.

- Metadata musí poskytnout entita (člověk, organizace, systém), který digitální dokumenty do repozitáře ukládá.
- Většina institucí ve světě, které se otázkou dlouhodobého uchovávání digitálních dokumentů seriózně zabývají, si tuto skutečnost uvědomuje. Jistá povědomost je i v NK ČR, jediným opatřením dosud bylo nasazení základní sady elementů ze standardů PREMIS a MIX do procesu digitalizace novodobých dokumentů v roce 2008. Obratem je až projekt NDK.
- Digitální objekt/-y a metadata dohromady tvoří tzv. Submission Information Package (SIP).
- Proces tvorby metadat do SIP balíčku musí být co nejvíce automatizován. To platí zvláště pro technická metadata, která vznikají automatickou extrakcí metadat pomocí externích nástrojů (např. JHOVE).
- Metadata SIP balíčku musí být v rozšířených standardech používaných v podobných projektech ve světě. Návrh profilu metadat pro NDK obsahuje nejrozšířenější standardy – METS pro celý balíček, PREMIS pro ochranná a technická metadata, MIX pro technická metadata digitálních obrazů, MODS a Dublin Core pro popisná metadata a ALTO XML pro vyjádření OCR na digitálních obrazech.
- Všechny dostupné a v disertační práci popisované SW systémy pro správu digitálních dat tyto typy metadat ve formě konkrétních metadatových standardů podporují, ať již jde o SW komerční (Rosetta – dříve DPS od firmy Ex Libris; systém SDB od firmy Tessella), nebo Open source (DSpace, Fedora, Archivematica aj.).
- SIP balíček obsahuje pouze základní sadu metadat, v rámci repozitáře uchovávaný archivní balíček (AIP) by měl obsahovat do jisté míry vyčerpávající metadatový popis (tj. všechny druhy metadat), která se v repozitáři k původním metadatům ze SIP balíčku doplňují. To jak velkou množinu metadat bude AIP obsahovat, musí opět určit instituce podle svých potřeb a možností.
- Z potřeby maximální informace doprovázející digitální objekt již z procesu digitalizace ve formě metadat vychází i návrh nového profilu metadat pro NDK, který je součástí disertační práce. Návrh profilu metadat pro digitalizaci v projektu NDK obsahuje technická, administrativní a ochranná metadata, která budou vznikat v procesu digitalizace (od roku 2012).

Teze 10.

Metadatový kontejnerový standard METS[3] (Metadata Encoding and Transmission Standard) se stal v oblasti metadat pro digitální objekty standardem a je ideální pro použití i v digitalizaci projektu *Národní digitální knihovna*.

- 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. Je to standard zamýšlený pro uložení metadat

[3] http://www.loc.gov/standards/mets/

elektronických textů, obrázků, videa, zvuku apod. Je strukturou pro vložení všech relevantních typů metadat:

- **popisných** – informace o intelektuálním obsahu objektu, podobně jako standardní záznam v katalogu; umožňuje uživateli digitální knihovny objekt najít a odhadnout jeho relevanci;
- **administrativních** – informace nutné pro správce digitální sbírky ke správě objektu, obsahuje informaci o intelektuálním vlastnictví, technické informace o objektu a souborech, které zahrnuje;
- **strukturálních** – informace o tom, jak jednotlivé části, které tvoří objekt, souvisí jedna s druhou, včetně pořadí v jakém mají být prezentovány uživateli.

- Standard METS umožňuje vložit tato metadata přímo do své struktury, nebo do externích souborů, na které je z METS odkazováno. METS je navržen přímo pro metadata digitálních knihoven [GARTNER, Richard. 2002].
- Vysvětleny jsou všechny části standardu, jeho použití obecně a vhodná implementace pro digitalizaci projektu NDK. Důraz je kladen na ty jeho části, které v sobě nesou ochranná metadata. Návrh využití standardu METS byl konzultován s odborníky z národních knihoven Norska, Finska a Nizozemí, které mají projekty masové digitalizace podobné NDK.

Teze 11.

Nejdůležitější částí standardu METS z pohledu dlouhodobého uchování digitálních objektů je část <amdSec>.

- Sekce administrativních **metadat „amdSec“** má sama další čtyři části:
  - <techMD> – technická a ochranná metadata;
  - <rightsMD> – administrativní případně legislativní práva k objektům;
  - <sourceMD> – popis původce údajů obsažených v METS dokumentu;
  - <digiprovMD> – metadata spojená s digitálními zdroji.
- Část <amdSec> je částí, do které se dle specifikace METS vkládají obecně administrativní metadata. V realitě jsou do ní vkládána i metadata technická a ochranná. Forma plnění této částí není ve vlastním standardu METS specifikována. Tyto sekce tedy musejí být „plněny“ již hotovým metadatovým popisem v určitém standardu nebo i standardech.
- Disertační práce, a návrh metadat pro NDK samotný, zohledňují tyto části a specifikují i metadatové standardy pro jejich plnění. Návrh profilu metadat pro NDK obsahuje konkrétní pravidla plnění jednotlivých podčástí <amdSec> METS standardu.

Teze 12.

Standard PREMIS je klíčový pro dlouhodobou ochranu digitálních dat.

- PREMIS[4] je metadatový standard vyvinutý s cílem vytvoření jednoduše použitelné sady základních elementů metadat pro logickou ochranu digitálních dokumentů, která by byla široce implementovatelná v komunitě zabývající se dlouhodobou ochranou digitálních dat.
- Ochranná metadata jsou informace podporující a dokumentující proces ochrany digitálních dokumentů: provenienci, autenticitu, ochranné aktivity, technické prostředí, management práv.
- Standard PREMIS lze použít samostatně, ovšem nejčastěji je zapouzdřen do kontejneru METS.

[4] http://www.loc.gov/standards/premis/

- PREMIS Datový slovník staví na referenčním modelu OAIS, který poskytuje koncepční základy specifikací druhů informačních objektů a balíčků pro archivované objekty a také strukturu s nimi souvisejících metadat [OCLC/RLG PREMIS WORKING GROUP. 2005, s. ix]. PREMIS a METS jsou vyvíjeny s ohledem na jednoduché vzájemné využití (záznam PREMIS vložený v záznamu METS).
- Vysvětlen je také datový model standardu PREMIS, který byl vyvinut pro vyjádření logické organizace metadatových elementů a entit PREMIS.
- Metadata ve standardu PREMIS musí vznikat již v digitalizaci a to pro jednotlivé části standardu, tedy PREMIS Object, PREMIS Event a PREMIS Agent. Všechny tyto části jsou součástí návrhu metadatového profilu pro digitalizaci projektu NDK. PREMIS bude zapouzdřen v záznamu METS, v jeho části <amdSec>.

5. Audit a certifikace digitálních repozitářů

V obecné rovině se nejčastěji počítá s tím, že informace budou v digitálním repozitáři uloženy velmi dlouho, přičemž budou stále čitelné, tj. použitelné. Vyhovět hlavnímu cíli většiny repozitářů není v době překotného vývoje technologií vůbec jednoduché. Technologie se vyvíjejí obrovským tempem, což přináší nejen zastarávání HW, ale hlavně zastarávání SW (viz výše). Pracovní skupina sestavená z odborníků organizací The Commission on Preservation and Access a The Research Libraries Group definovala digitální archivy jako: „*repozitáře digitálních informací, které jsou kolektivně odpovědné za zabezpečení, skrz uplatnění různých strategií migrace, integritu a dlouhodobou dostupnost národního sociálního, ekonomického, kulturního a intelektuálního dědictví ztělesněném v digitální formě.*“ [GARRETT a WATERS, 1996, s. 9-10] V Německu v rámci programu NESTOR[5] vznikla následující 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, 2009, s. 2]

V souvislosti s uložením a odpovídající kvalitou procesů v repozitáři se mluví o tzv. důvěryhodných repozitářích. Výraz „důvěryhodný“ označuje schopnost repozitáře zachovat digitální dokumenty v dlouhodobém horizontu přístupné a použitelné. Tedy repozitář musí být navenek důvěryhodný, což se netýká pouze technického řešení, právě naopak. Důvěryhodný repozitář musí prokázat, že provádí standardní procesy, má k nim odpovídající dokumentaci, disponuje odborným personálem, má zajištěné financování do budoucna, strategii ochrany dat a je schopen zajistit a prokázat integritu a autenticitu digitálních objektů, které uživateli poskytuje. Podstatná je také schopnost prokázat, že nebyla narušena jejich autenticita digitálních objektů (nikdo je záměrně neměnil), musí být jasný jejich původ, určení apod. Na objektech provedené zásahy, migrace apod. dokumentují administrativní metadata. Důvěryhodnost by měla být podložena a zaštítěna externí autoritou – auditorem, který ověří, že všechna tato hlediska nezbytná pro dlouhodobé uchování dokumentů jsou splněna.

Lze definovat deset základních principů důvěryhodnosti, které jsou jakýmsi průnikem všech dostupných certifikačních nástrojů a metodik. Lze je považovat za desatero repozitáře, který chce být považován za důvěryhodný. Takový archiv [CENTER FOR RESEARCH LIBRARIES, 2007]:

1. se musí zavázat k neustálému opatrování/správě digitálních objektů pro určitou cílovou komunitu.
2. musí prokázat svou životaschopnost/způsobilost (včetně financování, personálních otázek, struktury, procesů), aby dostál stanovenému závazku.

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

3. si musí osvojit a dodržovat potřebná smluvní a zákonná práva 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, jako 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 záchranných prací.
10. musí mít technickou infrastrukturu adekvátní pro účel neustálé údržby digitálních objektů.

Teze 13.

Vybudování důvěryhodného digitálního repozitáře a jeho certifikace jsou pro hodnotnou trvalou ochranu a zpřístupnění digitálních dokumentů nezbytné.

- Digitální repozitář v národní instituci (národní knihovna, národní archiv apod.) by měl svou důvěryhodnost prokázat absolvováním interního a externího certifikačního procesu. Předpokladem je připravenost, implementace referenčního rámce OAIS, vytváření dostatečných metadat.
- Důvěryhodný digitální repozitář musí prokázat schopnost provádět logickou dlouhodobou ochranu standardním způsobem. Tj. měl by mj. zajistit integritu a autenticitu dokumentů, ošetřit migraci (případně emulaci), ochranu proti neoprávněnému přístupu, poškození atd. K zajištění a uchování údajů o autenticitě i integritě napomáhají ochranná metadata.
- Z hlediska instituce snažící se v konečném výsledku o externí certifikaci, je vhodné absolvovat napřed interní „self“ audit, který pomůže odhalit konkrétní rizika a lze jej periodicky opakovat. Poté může absolvovat audit externí (placený), který může být díky internímu auditu kratší a také levnější.
- Podrobněji je rozebrán vnitřní audit pomocí nástroje DRAMBORA[6], který je nejvhodnější pro použití v českých paměťových institucích, kde se již používá.
- Disertační práce dále popisuje následující dokumenty a nástroje zabývající se kritérii hodnocení důvěryhodnosti:
    - Trustworthy Repositories Audit & Certification : Criteria and Checklist (OCLC, CRL, 2007), známý pod zkratkou TRAC;
    - NESTOR Criteria Catalogue.

[6] http://www.repositoryaudit.eu/

6. Závěr

Disertační práce prokázala platnost předložených tezí. Praktickou součástí práce je rozsáhlý návrh nového aplikačního profilu metadat, který bude využit od roku 2012 v digitalizaci projektu Národní digitální knihovna. Disertační práce mj. ukazuje, že v prostředí českých knihoven a archivů nebyla potřeba tvorby nových typů metadat zcela jednoznačně reflektována. V NK ČR vznikla první specifikace technických a ochranných metadat až v roce 2008 a to ve velmi základní podobě. 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. 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 disertační práce. 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. Návrh profilů, které jsou praktickou částí disertační práce, vznikaly intenzivně během let 2009-2011. Obsahují vedle popisných také potřebná metadata administrativní, ochranná, strukturální a technická. Oba návrhy jsou záměrně flexibilní a postavené pouze na schématech metadat, která jsou ve své oblasti ve světě běžně používanými standardy. Díky tomu lze metadata sdílet, vytvářet a upravovat v mnoha běžně dostupných nástrojích, které s konkrétními standardy pracují apod. 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[7], ale i v projektech jiných, jak tomu ostatně je v projektu digitalizace článků periodik ANL+.

[7] 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).

7. Přehled použitých informačních zdrojů pro teze disertační práce

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

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

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

GARTNER, Richard. 2002. *METS: Metadata Encoding and Transmission Standard* [online]. JISC 2002 [cit. 10-03-2012]. 12 s. Dostupné z: http://www.jisc.ac.uk/uploaded_documents/tsw_02-05.pdf

GIARETTA, David. **[1998].** ISO/CCSDS Open Archival Information System (OAIS) Reference Model [prezentace online]. [cit. 10-03-2012]. Dostupné z: http://ce.sharif.ir/courses/87-88/1/ce448/resources/root/EIAP-methodologies/OAIS1.ppt

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

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

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. 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

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