Univerzita Karlova v Praze
Filozofická fakulta

Ústav informačních studií - studia nových médií

Diplomová práce

Bc. Jaroslav Kvasnica

Dlouhodobé uchování webového obsahu

Long-term Preservation of Web Content

2016

Vedoucí práce: PhDr. Mgr. Jan Pokorný, Ph.D.

Prohlášení:

Prohlašuji, že jsem diplomovou 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, dne 26. července 2016

……

Jaroslav Kvasnica

Klíčová slova:

webová archivace; webový obsah; dlouhodobá archivace digitálních dokumentů; migrace; emulace; metadata; Národní knihovna České republiky

Keywords:

web archiving; web content; long-term preservation of digital information; migration; emulation; metadata; National Library of the Czech Republic

Abstrakt:

Tato práce řeší problematiku dlouhodobé ochrany digitálních dokumentů, konkrétně webových stránek. Cílem práce je tuto problematiku vysvětlit, vymezit rozdíly mezi různými přístupy k ní, popsat možnosti dlouhodobého uchování webového obsahu, jako je migrace a emulace nebo vysvětlit jaká jsou rizika a výzvy při zavádění těchto strategií.

Na základě této práce je možné si udělat představu o tom, jaké nové problémy přináší snaha o dlouhodobou ochranu webového obsahu a jaká jejich řešení jsou v současné době dostupná. Zároveň práce přináší pohled, jakým způsobem k problematice přistupují některé významné zahraničních institucí.

Hlavním výsledkem této práce je podrobná analýza strategie dlouhodobé digitální ochrany Národní knihovny České republiky, která je jedinou institucí zabývající se v takovém rozsahu ochranou českého webu. V práci je podrobně popsán proces přípravy dat, proces vytváření metadat a proces uložení do LTP úložiště NK ČR, včetně příkladů a jejich vysvětlení. V závěru práce je přiblíženo, jako další kroky český webový archiv čekají, aby byl schopný svá data dlouhodobě ochránit.

Abstract:

This work describes the long term preservation of digital documents, particularly websites. The aim of this work is to give an explanation of the long term preservation, to define the differences between various approaches and to describe long term preservation of web content possibilities such as migration or emulation. It also explains risks and challenges of these strategies. It discusses new problems which the long term preservation aim leads to. It also describes possible solutions as well as it describes the situation in selected significant foreign institutions.

The main aim of this work is detailed analysis of long term preservation strategy in theNational Library of the Czech Republic, which is the only institution engaged in the preservation of Czech web. The process of data preparation, metadata creation and data storing in the long term repository of the Czech National Library is thoroughly described, including examples and their explanation. Future actions of long term preservation in the Czech Web Archive are articulated in the conclusion.

Obsah

Seznam použitých zkratek

AIP	Archival Information Package
AJAX	Asynchronous JavaScript and XML
ALA	American Library Association
ALTO	Analyzed Layout and Text Object
API	Application Programming Interface
ARC	ARChive
BNF	Bibliothèque nationale de France
ČSN	Česká technická norma
DC	Dublin Core
DIP	Dissemination Information Package
DNS	Domain Name System
GB	Gigabyte
HTML	Hypertext Markup Language
IA	Internet Archive
IIPC	Internation Internet Preservation Consortium
IP	Internet Protocol
ISO	International Organization for Standardization
JPEG	Joint Photographic Experts Group
LTP	Long-term preservation
MB	Megabyte
MD5	Message Digest 5 Algorithm
METS	Metadata Encoding and Transmission Standard
MODS	Metadata Object Description Schema
NDK	Národní digitální knihovna
NISO	National Information Standards Organization
NK ČR	Národní knihovna České republiky
OAI-PHM	Open Archives Initiative - Protocol for Metadata Harvesting
OAIS	Open Archival Information System
OCR	Optical Character Recognition
PDF	Portable Document Format
PHP	Hypertext Preprocessor
PREMIS	Preservation Metadata Maintenance Activity

PSP	Producer Submission Package
RDF	Resource Description Framework
SIP	Submission Information Package
SPAR	Scalable Preservation and Archiving Repository
SW	Software
TB	Terabyte
TLD	Top-level domain
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
URN:NBN	Uniform Resource Name: National Bibliography Number
UUID	Universally unique identifier
VM	Virtual Machine
WA	Web Archive, webový archiv
WARC	Web ARChive
WWW	World Wide Web
XML	Extensible Markup Language

1. Úvod

Diplomová práce je věnovaná tématu dlouhodobého uchování webového obsahu. S rozšířením internetu se z něj stalo důležité médium, ve kterém se lidstvo v moderní době přirozeně kulturně projevuje. Veškerý webový obsah se tím stává nezastupitelným kulturním artefaktem, u kterého vzniká přirozená potřeba jej uchovat pro budoucnost. Archivace tak rozsáhlé lidské činnosti hraje důležitou roli, stejně jako u ostatních kulturních artefaktů v minulosti, při snaze o sebe-porozumění lidstva, formování vědomí společnosti a budování budoucnosti. World Wide Web (dále jen web) přinesl nejen nové možnosti lidské realizace, ale také nové výzvy pro knihovny a jiné paměťové instituce. Web je médium natolik odlišné od veškeré dosavadní lidské publikační činnosti, že je nutné radikálně upravit tradiční archivační metody.

Autor diplomové práce si téma dlouhodobé ochrany webového obsahu zvolil, protože v současné době vede Oddělení archivace webu v Národní knihovně České republiky (dále jen NK ČR), kde mimo jiné připravoval koncepci a koordinoval realizaci uložení dat z digitálního archivu českých webových zdrojů do úložiště dlouhodobé ochrany digitálních dokumentů.

Diplomová práce je rozdělena do čtyř částí, ve kterých je postupně probrána problematika dlouhodobého uložení webového obsahu. Tato problematika dlouhodobé ochrany dat z webu je velmi komplexní, interdisciplinární a přináší nové problémy a výzvy. Je to dáno zejména enormním objemem, dynamikou a celkovou komplexitou dat. Data vznikají a zanikají nahodile, internet není nijak centrálně řízený žádnou autoritou a data nejsou nijak standardizovaná, ani nemají garantovanou životnost. Data jsou tedy determinována svým původem, objevuje se v nich obrovské množství souborových formátů, jsou mezi sebou nahodile (z pohledu ochrany dat) propojena pomocí hypertextových odkazů. Vše se děje bez jakékoli kontroly nebo možné predikce. Zároveň se ale webové archivy stávají nepostradatelným nástrojem pro badatele, kteří se zabývají studiem moderní historie, a jejich důležitost bude narůstat spolu s narůstajícím stářím internetu.

V úvodu této práce je nutné uvést, že v současné době se vždy budeme pohybovat na poli spekulací jak se správně zachovat, abychom webová data ochránili. Je možné, a vysoce pravděpodobné, že za deset, dvacet nebo padesát let se veškeré strategie ochrany budou muset měnit a přizpůsobit technologickému vývoji, který s sebou budoucnost

přinese, protože úkolem dlouhodobé ochrany je uchovat data v řádech stovek let. I přes všechny překážky je nutné veškerý web archivovat a zajistit mu dlouhodobou ochranu, protože webový obsah nelze archivovat retrospektivně. Lze získat pouze jeho aktuální obraz.

První část diplomové práce je věnovaná celkovému úvodu do problematiky dlouhodobé ochrany. Jsou zde vysvětleny základní pojmy vztahující se k dlouhodobému uložení webového obsahu. Zejména pak půjde o vymezení pojmu webový obsah a to, jak se takový obsah liší od jiných typů dat. Úvodní část obsahuje stručný popis, jakým způsobem se data z webu získávají. Je naznačen základní rozdíl mezi dlouhodobou archivací běžných, tzn. jasně definovaných digitální objektů (např. digitalizovaných monografií nebo periodik) a webovým obsahem. Tato část je rovněž věnována normě OAIS, která je standardem pro budování repozitářů pro dlouhodobou ochranu. Nepůjde o kompletní výklad standardu, ale o popis základních principů, jakým způsobem fungují repozitáře digitální ochrany, jak k datům přistupují a jaká specifika s sebou nese ukládání dat z webu.

Druhá část diplomové práce přináší již hlubší vhled do problematiky na konceptuální úrovni. V první řadě se věnuje budování strategie digitální ochrany, která je základním nástrojem pro její realizaci. Schopnost zajistit ochranu dat v repozitáři na dlouhodobé úrovni je závislá na vypracování podrobné strategie ochrany. Tato strategie zaručuje, že repozitář umí reagovat na vývoj technologií, softwaru a hardwaru, ale také na nově vznikající potřeby uživatelů, pro které jsou data určena. Bez této strategie není možné zajistit, že data uložená v repozitáři budou využitelná v budoucnosti. Kroky, které vedou k vybudování takové strategie, se bude zabývat právě druhá část diplomové práce. Obě témata spojuje problematika metadat, které je také probíraná podrobněji v druhé části práce. Stěžejním tématem jsou konkrétní ochranná opatření, která poznatky z předchozí části přivádějí do praxe. Jedná se o migraci, emulaci a kombinaci obojího. V případě webového obsahu zvolení správné strategie představuje nelehký úkol, a to právě z důvodu komplexnosti dat.

Ve třetí části diplomové práce je popsána dlouhodobá archivace webového obsahu více z praktického hlediska. Věnuje se konkrétním institucím (Rakouské národní knihovně a Britské národní knihovně), které již nějakým způsobem s webovým obsahem pracují a snaží se jej dlouhodobě uchovávat se zaměřením hlavně na jejich softwarové

aplikace, metodiku a strategie. Součástí třetí části je i problematika dlouhodobého uložení tzv. webu 2.0, který s sebou nese jistá specifika. Tato specifika se projevují zejména v případě ukládání a zpřístupňování webových stránek. Odlišné přístupy budou demonstrovány na konkrétních projektech probíhajících ve světě.

Závěrečná část se věnuje dlouhodobé ochraně českého webového archivu (Webarchiv, dříve WebArchiv) NK ČR, který má na starosti archivaci českého webu. Tato data se ukládají do nově vybudovaného úložiště dlouhodobé ochrany, které vzniklo v NK ČR. Poslední část bude věnována kompletní cestě dat z dočasného úložiště Webarchivu až po uložení do dlouhodobého úložiště. Data prochází několika fázemi a stupni strukturalizace, validace a metadatového popisu. Veškeré tyto procesy jsou popsány ve třetí části, a to včetně příkladů z praxe. Součástí je také popis dalších kroků, které NK ČR nevyhnutelně čekají, aby dokázala svá data dlouhodobě ochránit.

2. Dlouhodobé uložení webového obsahu

V následující kapitole a jednotlivých podkapitolách jsou vysvětleny základní pojmy, které jsou důležité pro další části práce – zejména pak pro závěrečnou část popisující současnou praxi. Nejprve je třeba definovat, co je vlastně myšleno slovním spojením webový obsah.

2.1. Webový obsah v prostředí webových archivů

Obecně lze za webový obsah označit vše, co se nějakým způsobem nachází na internetu, a tím mohou být myšleny nejen veřejně dostupné webové stránky, ale také např. databáze, skripty, případně veškerý deep web[1]. Pro potřeby této práce je webovým obsahem myšlen pouze povrchový web[2]. Webový obsah je pak v rámci oboru webové archivace, velmi zjednodušeně, obsah, který byl stažen z internetu a uložen v nějakém archivu, mluvíme pak o archivním webovém obsahu nebo o archivních kopiích webového obsahu.

Akvizice takového obsahu se nazývá archivace webu (ang. web archiving) a její specifikum spočívá ve dvou bodech: „*Zaprvé – předmětem její činnosti jsou výhradně dokumenty, které jsou volně zpřístupňované internetovou sítí. Zadruhé – akvizice těchto digitálních dokumentů se provádí formou jejich automatického získávání (…)*“[3] Z toho plyne, že webový obsah nejenže pochází z internetu, ale jedná se o obsah, který je běžně volně přístupný všem uživatelům.

Archivaci webu lze rozdělit na dva základní druhy a to na institucionální a personální. Do personální spadá archivace webu pro osobní účely, tedy archivace na straně vlastníka webového serveru. Nejčastěji se jedná o archivaci, kterou provádí firmy v rámci archivace svých interních materiálů, do kterých je zahrnuta i webová prezentace firmy nebo její intranet. Institucionální archivace webu provádějí veřejné instituce – nejčastěji národní knihovny (Evropa) nebo univerzity (USA), které se snaží zachovat tu část internetu, která spadá zaměřením do jejich fondu, za účelem uchování kulturního dědictví a zpřístupnění archivovaného obsahu svým uživatelům.

---

[1] Deep web jsou webové stránky, které nejsou indexované pomocí vyhledávačů

[2] Povrchový web je část internetu, která je indexovaná a dostupná běžným vyhledávačem.

[3] CUBR, Ladislav. *Dlouhodobá ochrana digitálních dokumentů*. 1. vyd. Praha: Národní knihovna České republiky, 2010, s. 35. ISBN 978-80-7050-588-5.

Oba tyto druhy archivace webu mají svá technická specifika. Existují tři základní přístupy, jak technicky archivovat web. Archivace webu na straně klienta (ang. client-side web archiving), archivace webu na straně serveru (ang. server-side web archiving) a archivace webu založená na transakcích (ang. transaction-based web archiving).[4] Archivace na straně serveru je přímé kopírování souborů ze serveru a archivace založená na transakcích „*sleduje, jak se skuteční uživatelé pohybují v rámci stránek, a zaznamenává jednotlivé transakce na serveru včetně archivace doručeného obsahu.*“[5] Oba tyto přístupy vyžadují přístupové údaje k webovým stránkám nebo serveru. Není tedy možné je praktikovat ve velkém měřítku, neboť potřebují přímou spolupráci s vlastníkem webové stránky při její archivaci. Tyto přístupy jsou využívány zejména pro personální archivaci webu.

Archivace na straně klienta spočívá ve stahování stránek tak, jak je vidí běžný anonymní uživatel při prohlížení. Není stahován obsah pro jehož zobrazení je třeba autorizace nebo autentizace, a nejsou stahovány ani zdrojové kódy, ke kterým se není možné dostat bez přímého spojení na server (např. zdrojové kódy v jazyku PHP jsou prohlížečem interpretovány a uživatel je vidí jako HTML). Tento technický přístup se používá při institucionální archivaci webu, protože je možné jej použít ve velkém měřítku.

2.1.1. Large-scale web archiving

Téměř veškerá institucionální archivace webu probíhá ve velkém měřítku (ang. large-scale web archiving). To znamená, že webový archiv se snaží archivovat tak velkou část internetu, že není možné kontrolovat akvizici, ochranu, zpřístupnění ani kvalitu jednotlivých webových stránek pouze lidskými silami. V takovém případě zde nastupují automatizované procesy.

To, jak velkou část internetu archiv pokrývá, je dáno jeho zaměřením. Jak bylo zmíněno výše, webové archivy jsou velmi často součástí národních knihoven, a proto je lze označit za národní webové archivy. Tyto archivy se pak zpravidla zaměřují

---

[4] Web Archiving Guidance. *The National Archives: The UK government's official archive* [online]. 2011 [cit. 2015-09-04]. Dostupné z: http://www.nationalarchives.gov.uk/documents/information-management/web-archiving-guidance.pdf

[5] COUFAL, Libor. Web po 20 letech: co z něj zbude pro budoucí generace?. *Knihovna* [online]. 2009, roč. 20, č. 2, s. 17-32 [cit. 2015-09-04]. Dostupné z: http://knihovna.nkp.cz/knihovna92/0902097.htm

na produkci svého národa. Například český webový archiv se zaměřuje na sběr: „*bohemikálních dokumentů zveřejněných v prostředí sítě Internet. (…) Provádí se jednak kompletní plošná archivace, tj. automatický sběr ‚celého' českého webu, souběžně však probíhá i výběrová archivace (nejzajímavějších webových zdrojů vybraných na základě selekčních kritérií a tematické archivace (zaměřené na určité aktuální téma, např. volby, povodně apod.)*"[6] V případě univerzit se jejich zaměření udává na základě vyučovaných oborů a zájmů studentů či spřízněných výzkumníků.

Na příkladu českého archivu lze ilustrovat to, o jak velké množství dat se jedná. Český web nejenže zahrnuje všechny domény .cz, ale také se jedná o zdroje, které splňují alespoň jednu z těchto podmínek: „*Byly vydány na území České republiky; jsou v českém jazyce (bez ohledu na místo vydání); autor je Čech; předmět se týká České republiky nebo českého národa.*"[7] V současné době se jedná o přírůstek řádově v desítkách terabajtů ročně. Pro sklízení takového množství webového obsahu je potřeba specializovaný software, který se nazývá sklízecí robot (ang. crawler nebo harvester). „*V současné době je jedním z nejvíce rozšířených open-source program Heritrix, který je využíván i v projektu WebArchiv. Heritrix, vyvíjený v rámci IIPC[8] pod vedením americké organizace Internet Archive, funguje na podobném principu jako roboti internetových vyhledávačů.*"[9]

Pro potřeby Heritrixu a potřeby webových archivů obecně byl vyvinut speciální souborový formát, který se nazývá ARC (ARChive), ze kterého pak vznikla jeho vylepšená verze WARC (Web ARChive).

2.1.2. Souborové formáty ARC a WARC

ARC je archivní kontejnerový formát umožňující ukládání webového obsahu, který je přirozeně v různých souborových formátech, do jednoho souboru. ARC je velmi jednoduchý souborový formát, který vznikl za účelem ulehčit manipulaci s velkým

---

[6] BROKEŠ, Adam. Projekt WebArchiv: archiv českého webu. *Zpravodaj ÚVT MU: bulletin pro zájemce o výpočetní techniku na Masarykově univerzitě* [online]. Brno: Ústav výpočení techniky MU, 2008, XVIII, č. 4, s. 10-13 [cit. 2015-09-04]. Dostupné z: http://www.ics.muni.cz/zpravodaj/articles/578.html

[7] CELBOVÁ, Ludmila. *Archivace webu*. 1. vyd. Praha: Národní knihovna ČR, 2008, 45 s. ISBN 978-80-7050-562-5.

[8] International Internet Preservation Consortium je mezinárodní organizace sdružující knihovny a jiné organizace zřízené za účelem koordinace úsilí o zachování přístupu na webové stránky pro budoucnost.

[9] GRUBER, Lukáš, Tomáš SÍBEK a Libor COUFAL. Archivace webových stránek v českém prostředí aneb Jak funguje WebArchiv. Čtenář [online]. Kladno: Středočeská vědecká knihovna v Kladně, 2009, **2009, 61**(5/2009) [cit. 2016-07-26]. ISSN 1805-4064. Dostupné z: http://ctenar.svkkl.cz/clanky/2009-roc-61/05-2009/tema-archivace-webovych-stranek-v-ceskem-prostredi-aneb-jak-funguje-webarchiv-58-393.htm

objemem souborů a jejich agregováním. V ARCu se nachází jen velmi málo metadat a obsah kontejneru se skládá ze dvou částí. Z hlavičky, která obsahuje jedinečné jméno souboru, IP adresu, čas pořízení souboru, typ a přesnou velikost – obzvláště důležitou částí hlavičky je také formát záhlaví následujících záznamů a z těla, které obsahuje URL záznam, záhlaví a samotná data, případně další metadata.[10] To znamená, že každý jednotlivý soubor má přidělenou hlavičku s metadaty.

WARC je nový archivní formát, který je vylepšenou verzí zastaralého ARCu. WARC přináší řadu vylepšení, zejména pokud se jedná o metadatový popis jednotlivých souborů v něm uložených. Formát WARC je jednoduché zřetězení jednoho nebo více tzv. WARC záznamů. První záznam obvykle popisuje záznamy, které budou následovat. Obecně platí, že záznamy jsou dvojího typu. Prvním typem jsou záznamy pro obsah, které jsou přímým důsledkem pokusu o vyhledávání informací – webové stránky, vložené obrázky, informace, přesměrování URL, výsledky DNS hostname vyhledávání, samostatné soubory. Druhým typem jsou záznamy se syntetizovanými daty (metadata), které poskytují další informace o archivovaném obsahu.[11] Každý jednotlivý soubor uložený v kontejneru je opatřený hlavičkou, která obsahuje jeho technický popis a také cestu, kudy pokračovat při zpětné rekonstrukci webové stránky.

Při zpětné rekonstrukci webových stránek je opět nutné využít specializovaný software, v dnešní době je nejpoužívanější Wayback Machine. Wayback Machine je open-source software vyvíjený pod hlavičkou Internet Archive, který slouží k zpřístupnění webových archivů uživatelům. Tento software pracuje právě s formáty ARC i WARC. Spojuje webový obsah z kontejnerů zpět do webových stránek a umožňuje uživateli jejich procházení napříč časem. S kontejnerovými formáty souvisí i problém roztříštěnosti jednotlivých fragmentů webových stránek do různých kontejnerů. Záleží na nastavení sklízecích robotů, ale většinou se stránky sklízejí paralelně a kontejnery musejí mít definovanou maximální velikost. Proto jednotlivé webové stránky mohou být uloženy ve více na sebe nenavazujících kontejnerech. Z toho důvodu není nadále možné počítat s kontejnerem jako s hlavní logickou jednotkou pro práci s daty uloženými ve webovém archivu.

---

[10] ZACH, Michael. Celosvětový Archiv Internetu a jeho role v získávání, uchování a zpřístupňování webových zdrojů. Praha, 2007. Bakalářská práce. Univerzita Karlova v Praze. Vedoucí práce PhDr. Eva Bratková.

[11] Web Archiving: Issues and Methods. MASANÈS, Julien. *Web archiving*. New York: Springer, c2006, s. 2-53. ISBN 3540233385-.

2.1.3. Digitální objekty

Jakmile se webový obsah uloží do archivu, je nutné se na něj přestat dívat jako na soubory nebo fragmenty webových stránek agregovaných v kontejnerových formátech. Při práci s daty v archivu již nestačí pracovat jen s kontejnery, protože při tak obrovském objemu dat není možné zachovat podrobný přehled o tom, co je v nich uloženo. A zároveň v dnešní době také nelze jednoduše pracovat s jednotlivými soubory uloženými v kontejneru, neboť na to nestačí výpočetní výkon.

Další specifikum, které webový obsah přináší, je jeho strukturovanost a jeho vzájemná propojenost, kterou ovšem není možné ze strany archivace webu nijak ovlivnit. Webové archivy se musí potýkat s obsahem, který je živě propojený, vzájemně prokládaný a je uložený v bohatých informačních strukturách, které jsou, stejně jako obsah, neustále vytvářeny miliony lidmi.[12] Proto je nutné pro práci s archivem definovat virtuální jednotku zvanou digitální objekt. Ta vznikne spojením dat a metadat popisující data.

Na digitální objekt je nahlíženo ze tří perspektiv: jako na fyzický objekt, logický objekt a konceptuální objekt.[13] Fyzická perspektiva se dívá na digitální objekt jako na shluk jedniček a nul uložených na fyzickém médiu. Logický digitální objekt reprezentuje digitální soubor, se kterým pracuje pro tento účel určený software. A konceptuální objekt je objekt, se kterým pracuje a rozumí mu uživatel. Při dlouhodobé ochraně digitálního objektu je nutné pracovat se všemi třemi perspektivami.

Referenční model Open Archival Information System (dále jen OAIS, česky Otevřený archivační informační systém) je zásadní ISO norma pro celý obor dlouhodobé archivace (podrobněji o normě v kap. 2.2.1) a jde s definicí digitálního objektu dále než výše citovaný Michael Day. Referenční model OAIS pracuje s informačním objektem, který je *„složen z datového objektu, který může být buď fyzický, nebo digitální, a z vysvětlujících informací, které umožňují data v úplnosti převést do podoby významově bohatších informací“*[14].

---

[12] Web Archiving: Issues and Methods. MASANÈS, Julien. *Web archiving*. New York: Springer, c2006, s. 2-53. ISBN 3540233385-.

[13] DAY, Michael. The Long-Term Preservation of Web Content. MASANÈS, Julien. *Web archiving*. New York: Springer, c2006, s. 177-199. ISBN 3540233385-.

[14] ČSN ISO 14721. Systémy pro přenos dat a informací z kosmického prostoru – Otevřený archivační informační systém – Referenční model. Praha: Úřad pro technickou normalizaci, metrologii a státní zkušebnictví, 2014.

Pro účely pokročilejší práce s digitálními objekty ve webových archivech vznikla rozšířená definice digitálního objektu, kdy je samotný digitální objekt rozšířen o časový údaj. Takový objekt se nazývá memento a je zapouzdřením stavu digitálního objektu v určitém čase.[15] Z toho vyplývá, že různá mementa stejného digitálního objektu se mohou vzájemně lišit dle toho, jak byl měněn jejich stav na živém webu a v jaký čas byla uložena do archivu. Koncept mementa se využívá pro plně distribuovaný systém, ve kterém různé verze digitálních objektů mohou být uloženy v různých archivech nebo na různých serverech. [16]

2.1.4. Shrnutí

Z výše uvedeného vyplývá, že přístupů a směrů k archivaci webu existuje velké množství a slovní spojení webový obsah tedy může nabývat více významů. Proto bylo nutné stručně vysvětlit základní pojmy, které se v této oblasti vyskytují. Účelem této práce není zabývat se archivací webu z pohledu akvizice zdrojů nebo jejich zpřístupňování, ale jejich dlouhodobou ochranou. Přestože ostatní problematiky s dlouhodobou ochranou souvisí a v práci se objevují, jsou vysvětleny jen stručně a ne příliš dopodrobna. Archivace webu je v dnešní době velmi obsáhlý obor, který se dynamicky rozvíjí.

Pro účely této práce je tedy webový obsah definován jako obsah, který je generován institucionálním sklízením ve velkém měřítku za účelem zachování kulturního dědictví napříč časem. Takový obsah je nejvíce komplexní a přináší spoustu problémů a výzev pro oblast dlouhodobé ochrany digitálních dokumentů.

---

[15] DE SOMPEL, Van. HTTP Framework for Time-Based Access to Resource States: Memento. IETF Tools: ETF-related tools, standalone or hosted on tools.ietf.org. [online]. USA, 2013 [cit. 2016-07-25]. Dostupné z: https://tools.ietf.org/html/rfc7089#section-1.1

[16] Tamtéž jako 15.

2.2. Dlouhodobá ochrana digitálních dokumentů

Dlouhodobá ochrana digitálních dokumentů nebo také dlouhodobé uložení, dlouhodobá archivace digitálních dokumentů (ang. long-term preservation, long-term archiving) je oblastní výzkumu a praxe, která má poslání uchovat digitální informace v čase ve formě přístupné koncovému uživateli. American Library Association (dále jen ALA) definovala dlouhodobou ochranu digitálních dokumentů jako: „*kombinaci politiky, strategie a opatření, které mají zajistit přístup k digitalizovaným a born-digital dokumentům bez ohledu na problémy s fyzickými nosiči a technologický vývoj. Cílem dlouhodobé ochrany digitálních dokumentů je zobrazení autentického obsahu v průběhu času.*“[17]

V definici ALA lze nalézt hned několik zásadních pojmů pro celou oblast dlouhodobé ochrany digitálních dokumentů. Pokud se jedná o základní dělení digitálních dokumentů, mohou to být buď tzv. born-digital anebo dokumenty digitalizované. Dokumenty, které jsou born-digital, byly vytvořeny a nadále s nimi bylo nakládáno pouze v digitální formě, např. digitální fotografie. V případě digitalizovaných dokumentů se jedná o dokumenty, které prvotně existovaly v analogové formě, např. naskenovaná papírová monografie. Co se týče webového obsahu, jeho drtivá většina je born-digital, přestože zde jsou k nalezení i digitalizované dokumenty.

Z tohoto rozdělení vychází Paul Conway, který ve svém článku rozlišuje digitalizaci pro ochranu (ang. digitalization for preservation) a digitální ochranu (ang. digital preservation). Zatímco digitalizace pro ochranu vytváří nové digitální dokumenty, jejichž hodnota je poté chráněna nezávisle na originální analogové kopii[18], digitální ochrana se zabývá ochranou dokumentů, při jejichž vzniku nefigurovala. Pokud se tedy nachází mezi webovým obsahem, který je archivován, dokumenty pocházející z digitalizace, tak na ně takto není možné nahlížet. Při dlouhodobé ochraně webového obsahu se tedy musí na všechny dokumenty dívat jako na born-digital obsah.

---

[17] z ang. “Digital preservation combines policies, strategies and actions to ensure access to reformatted and born digital content regardless of the challenges of media failure and technological change. The goal of digital preservation is the accurate rendering of authenticated content over time.”
Definitions of Digital Preservation. ALCTS: Association for Library Collections & Technical Services [online]. Washington, D.C., 2007 [cit. 2016-07-25]. Dostupné z: http://www.ala.org/alcts/resources/preserv/defdigpres0408

[18] CONWAY, Paul. Preservation in the Age of Google: Digitization, Digital Preservation, and Dilemmas. *Library Quarterly* [online]. Chicago: University of Chicago, 2010, vol. 80, no. 1, s. 61-79 [cit. 2015-09-16]. Dostupné z: http://deepblue.lib.umich.edu/bitstream/handle/2027.42/85223/J15%20Conway%20Preservation%20Age%20of%20Google%202010.pdf

Digitalizace pro ochranu a digitální ochrana spolu úzce souvisejí, ale základní normy, procesy, technologie, náklady a organizační problémy jsou zcela odlišné.[19]

Dalším zásadním pojmem z definice dlouhodobé ochrany od ALA je autenticita. Při dlouhodobé ochraně se dbá na autenticitu obsahu, tedy na pravost nebo hodnověrnost. „*Autenticita je obecně vlastnost, kterou získávají jakékoli reálie, u nichž není pochyb, že jsou skutečně tím, za co se vydávají.*“[20] Tím, že jednou z vlastností digitálních dokumentů je snadná možnost editace, nabývá autenticita na velké důležitosti pro oblast digitální ochrany. „*Obecně lze k těmto dokumentům přistupovat tak, že jejich autenticita je prokázána metadaty, tedy údaji o všech procesech. (…) digitální dokument je autentický, pokud byl vytvořen a zaslán oprávněnou osobou, v relevantním čase, a je tím, zač se vydává. (…) Společnost, tedy uživatelé, musí mít také samozřejmě důvěru v instituci archivu samotného.*“[21]

Hned na začátku definice se objevují pojmy politika, strategie a opatření. Instituce, které se zabývají dlouhodobou ochranou, již z podstaty věci musejí mít svoje strategie formálně a závazně zakotvené v oficiálních dokumentech. Tyto strategie vznikají pro dlouhodobý výhled a nesmí na ně mít dopad fluktuace zaměstnanců, ani problémy s financováním archivu.

Dokumenty strategie dlouhodobé ochrany mají za cíl definovat povinnosti instituce, která spravuje úložiště dlouhodobé ochrany a také má poskytnout vodítko pro odpovědné zaměstnance při rozhodování a provádění činností, které mají přímý dopad na chráněná data.[22] V dokumentech musí být také popsaná strategie financování archivu, neboť se jedná o důležitý aspekt pro archiv. Zodpovědností instituce není jen ochránit data, ale také dlouhodobě udržet vlastní archiv v provozu.

---

[19] CONWAY, Paul. Preservation in the Age of Google: Digitization, Digital Preservation, and Dilemmas. *Library Quarterly* [online]. Chicago: University of Chicago, 2010, vol. 80, no. 1, s. 61-79 [cit. 2015-09-16]. Dostupné z: http://deepblue.lib.umich.edu/bitstream/handle/2027.42/85223/J15%20Conway%20Preservation%20Age%20of%20Google%202010.pdf

[20] DVOŘÁK, Tomáš. Uchovávání digitálních dokumentů se zachováním jejich autenticity. *Ústav informačních studií a knihovnictví: Jinonické informační pondělky* [online]. 2010 [cit. 2015-09-19]. Dostupné z: http://uisk.ff.cuni.cz/dwn/1003/14262cs_CZ_jip.pdf

[21] MELICHAR, Marek a Jan HUTAŘ. České paměťové instituce a digitální data: historický exkurz, současný stav a předpokládaný vývoj III. *Duha* [online]. 2014, roč. 28, č. 2 [cit. 2015-09-19]. Dostupné z: http://duha.mzk.cz/clanky/ceske-pametove-instituce-digitalni-data-historicky-exkurz-soucasny-stav-predpokladany-vyvoj-1

[22] Preservation policy. NATIONAL LIBRARY OF AUSTRALIA. *National Library of Australia* [online]. 2009 [cit. 2015-09-19]. Dostupné z: http://www.nla.gov.au/policy-and-planning/preservation-policy

2.2.1. Referenční model OAIS

Referenční model OAIS[23] je zkrácený název pro český překlad mezinárodní normy ISO 14721:2012, která popisuje otevřený archivační informační systém. Tento dokument se stal základní normou pro oblast dlouhodobé archivace. „*Lze jej aplikovat na jakýkoli archiv, jeho hlavní zaměření se však týká dlouhodobé ochrany digitálních dat. Model OAIS se stal společným terminologickým a konceptuálním rámcem všech projektů ochrany digitálních dat.*“ [24] Referenční model OAIS je natolik fundamentální dokument, že se pojmy, které zavádí, objevují napříč celou diplomovou prací. Stejně tak byl rámcem pro vytváření konceptu dlouhodobé ochrany webového obsahu v NK ČR, kterému je věnována poslední část této práce.

Jednou z částí normy OAIS je definice šesti základních povinností, které je nutné splnit, aby byly naplněny hlavní cíle dlouhodobé digitální archivace. Archiv podle OAIS musí:[25]

- Vyjednávat s tvůrci informací a přijímat od nich příslušné informace.
- Získávat možnost s poskytnutými informacemi dostatečně nakládat, aby bylo možné zajistit jejich dlouhodobou ochranu.
- Určit, které skupiny uživatelů by se měly stát určenými skupinami[26].
- Zajistit, aby informace určené k uchovávání byly pro určenou skupinu srozumitelné samy o sobě.
- Dodržovat zdokumentovaná pravidla a postupy, které zajistí, že informace budou chráněny.
- Zpřístupňovat uchovávané informace určené skupině.

Díky těmto povinnostem lze vidět, že OAIS norma se zabývá digitální ochranou velmi komplexně. A to jak od prvotního získávání dat, přes jejich ochranu až po finální zpřístupnění dat uživatelům.

---

[23] Celý název normy: Systémy pro přenos dat a informací z kosmického prostoru – Otevřený archivační informační systém – Referenční model

[24] LUKŠŮ, Alžběta. *Dlouhodobé uchovávání a zpřístupňování dokumentů zaznamenaných na optických discích*. Brno, 2010. Dostupné z: http://is.muni.cz/th/262809/ff_b/Bakalarska_prace_Alzbeta_Luksu.txt. Bakalářská práce. Masarykova univerzita.

[25] ČSN ISO 14721. Systémy pro přenos dat a informací z kosmického prostoru – Otevřený archivační informační systém – Referenční model. Praha: Úřad pro technickou normalizaci, metrologii a státní zkušebnictví, 2014.

[26] Určená skupina (ang. designated community) je identifikovaná skupina potenciálních uživatelů, kteří by měli rozumět zpřístupňovaným informacím.

Hned u prvního bodu lze narazit na problém při dlouhodobé archivaci webového obsahu, neboť vyjednávat dohodu s tvůrci informací není možné kvůli samotné povaze publikování informací na internetu. V tomto případě musí instituce, která provozuje webový archiv, převzít zodpovědnost na sebe a definovat sebe jako tvůrce informací, které vytváří automatizovaným sklízením webových zdrojů.

Do jisté míry je každá sklizeň unikátní a autorská, protože je determinována časem spuštění (internet se neustále proměňuje), kurátorskou prací (u výběrových a tematických sklizní), ale i vlastním technickým nastavením sklízecího robota (nastavení hloubky sklízení, reakce na chyby atd.). Sklizeň, *„představuje jeden ukončený proces stahování všech webových stránek podle určitých kritérií“*[27]. V tomto případě veškeré vyjednávání odpadá a musí zde nastoupit odborná diskuze, jak je s daty dále možné nakládat v rámci příslušného autorského zákona.

2.2.2. Dlouhodobá ochrana – fyzická

Dlouhodobá ochrana fyzická je základním stavebním kamenem pro celou oblast digitální ochrany. Nicméně pokud by zůstalo jen u ní, pak nejde o dlouhodobou ochranu, ale principiálně pouze o zálohu dat. Na této úrovni se řeší ochrana zejména z hlediska technologického, tedy na jaký hardware a jak archivní data ukládat. Známým problémem je zastarávání fyzických nosičů, které je v případě zničení nebo opotřebení pro data fatální. Pokud je poškozena část analogového nosiče, např. knihy, tak ostatní části zůstávají čitelné. Ale pokud je poškozen např. optický disk, tak zpravidla dochází ke ztrátě veškerého obsahu.

Na této úrovni dlouhodobé ochrany je zásadní volba nosiče a pak plán údržby, tedy vytváření kopií, kontrola životnosti nosičů, plán jejich výměny apod. Při volbě fyzických nosičů by měla být zohledněna tato kritéria: *„životnost nosiče, kapacita, použitelnost, zastarávání, náklady – pořizovací cena, náchylnost ke zničení“*[28].

---

[27] KVASNICA, Jaroslav a Rudolf KREIBICH. Formátová analýza sklizených dat v rámci projektu WebArchiv NK ČR. *ProInflow*. 2013, 5. ročník, 2. číslo. Dostupné z: http://pro.inflow.cz/formatova-analyza-sklizenych-dat-v-ramci-projektu-webarchiv-nk-cr

[28] BROWN, Adrian. Digital Preservation Guidance Note: Selecting Storage Media for Long-Term Preservation. THE UK GOVERNMENT'S OFFICIAL ARCHIVE. *The National Archives* [online]. 2008 [cit. 2015-09-30]. Dostupné z: http://www.nationalarchives.gov.uk/documents/selecting-storage-media.pdf

Dlouhodobá ochrana na fyzické úrovni nepředstavuje pro dlouhodobé uložení webového obsahu žádné výzvy navíc. Samozřejmě kromě většího objemu dat se technicky nijak neliší od jiného digitálního obsahu. Z těchto důvodů tato úroveň ochrany nebude v práci již nadále řešena.

2.2.3. Dlouhodobá ochrana logického a konceptuálního digitálního objektu

Na rozdíl od fyzické ochrany není dlouhodobá ochrana logického a konceptuálního digitálního objektu pouze problémem čistě technickým, ale přináší řadu nových problémů, které jdou nad rámec zastarávání a selhávání hardwaru. „*Podstatou je dosáhnout toho, aby dokument byl čitelný, použitelný i v budoucnu. V tomto smyslu je nutné provádět na dokumentech změny a ty zaznamenat (změnou je nejčastěji myšlena migrace do jiného formátu).*“[29] Kvůli vysoké rychlosti vývoje informačních technologií se u dlouhodobé ochrany objevují překážky, které Ladislav Cubr ve své knize ještě dělí do čtyř rovin: informační, systémové, institucionální a technologické[30] (do které patří problematika z předchozí kapitoly).

V rovině institucionální se jedná o překážky spojené s financováním archivu a také problémy spojené s lidskými zdroji nebo důvěryhodností repozitáře. Stejně jako u dlouhodobé ochrany fyzické, je tato problematika stejná jako u běžné dlouhodobé ochrany, proto v práci nebude dále hlouběji probírána.

U informační roviny se jedná o rizika spojená se souborovými formáty a jejich dokumentací, zejména rizika spojená se zastaráváním a podporou formátů, i rizika související s ohledně vlastnictvím, patentovými překážkami, ale také otevřeností dokumentace a specifikace souborových formátů. „*Na systémové rovině je třeba řešit dlouhodobou ochranu fyzické, logické a konceptuální vrstvy digitálního objektu, které jsou seskupeny v organizovaných sbírkách.*“[31] U systémové roviny se objevují problémy spojené s identifikací dokumentů, s digitálními právy nebo integritou. Je ale také třeba potýkat se se správou velkého počtu digitálních dokumentů. Dlouhodobá ochrana webového obsahu přináší v rovině informační a systémové zcela nové výzvy, a proto je jí věnována celá následující sekce.

---

[29] MELICHAR, Marek a Jan HUTAŘ. České paměťové instituce a digitální data: historický exkurz, současný stav a předpokládaný vývoj III. *Duha* [online]. 2014, roč. 28, č. 2 [cit. 2015-09-19]. Dostupné z: http://duha.mzk.cz/clanky/ceske-pametove-instituce-digitalni-data-historicky-exkurz-soucasny-stav-predpokladany-vyvoj-1

[30] CUBR, Ladislav. *Dlouhodobá ochrana digitálních dokumentů*. 1. vyd. Praha: Národní knihovna České republiky, 2010, 154 s. ISBN 978-80-7050-588-5.

[31] Tamtéž jako 29.

3. Možnosti dlouhodobého uchování webového obsahu

Třetí kapitola se věnuje již konkrétním problémům, výzvám a opatřením, kterým je nutné věnovat pozornost, pokud má být zachována dlouhodobá použitelnost webového obsahu.

3.1. Webové archivy a dlouhodobá ochrana

V úvodu třetí kapitoly je nejprve třeba určit, jaký je vztah mezi webovým archivem resp. oborem webové archivace (ang. web archiving), a dlouhodobou ochranou webového obsahu (ang. web preservation). Tyto dva termíny jsou totiž mezi sebou velmi často zaměňovány. Ve skutečnosti je mezi nimi velký rozdíl. Zatímco oblast dlouhodobé ochrany webového obsahu je podmnožina oboru dlouhodobé archivace digitálních dokumentů, oblast webové archivace je samostatný obor, který se zabývá několika procesy.

Procesy archivace webu:[32]

- Selekce: kurátoři vybírají webové zdroje, které jsou vhodné k archivaci.
- Akvizice, sklízení: činnost, kdy se pomocí softwaru vytváří kopie vybraných webových zdrojů a ukládají se do archivu.
- Zpřístupnění: webový archiv zpřístupňuje svůj obsah koncovým uživatelům.

Drtivá většina webových archivů deklaruje, že jejich úkolem je sklizená data dlouhodobě ochránit. Dlouhodobá ochrana webového obsahu se stará o to, aby data uložená v archivu mohla být v budoucnosti stále zpřístupňována koncovým uživatelům archivu. Obor dlouhodobé ochrany webového obsahu je velmi mladá oblast bádání, která zdaleka nemá vyřešené všechny své problémy a v dnešní době ještě není zodpovězena otázka, jak správně webový obsah dlouhodobě ochránit. Je to dáno zejména tím, že webové archivy se v minulých letech soustředily především na vývoj a optimalizaci samotné akvizice, a proto téma dlouhodobá ochrany bylo upozaděno.

[32] BRENDA AYALA, Reyes. Web Archiving Bibliography 2013. In: *UNT Digital Library* [online]. 28. 6. 2013 [cit. 2015-10-12]. Dostupné z: http://digital.library.unt.edu/ark:/67531/metadc172362/m1/1/

3.2. Rizika a výzvy

Dlouhodobá archivace webu přináší zcela nové výzvy v oblasti dlouhodobé archivace digitálního obsahu. Je to dáno specifičností webového obsahu. Jedním ze specifik je neustálý rychlý nárůst dat, není tedy v lidských silách uchovat vše a ani při kurátorském výběru není možné uchovat to důležité. S tím souvisí i životnost samotného obsahu, který se neustále vyvíjí, mizí a objevuje se, nebo dochází k jeho změnám, restrukturalizaci apod.[33]

Další rizika a výzvy souvisejí s neustálým vývojem webových technologií. Webové zdroje používají různé souborové formáty, programovací či skriptovací jazyky, některé webové zdroje používají technologie, které potřebují speciální plug-in, nebo fungují jen v určitém internetovém prohlížeči. Jiné webové zdroje využívají dynamické databáze pro uložení obsahu, další mohou být zabezpečené nebo být součástí hlubokého webu. Toto vše přispívá k tomu, že dlouhodobá ochrana webového obsahu je z technologického hlediska velmi náročná.

Z technického pohledu je zásadním rozdílem, který přináší dlouhodobá ochrana webového obsahu to, že webový obsah sám o sobě je velmi různorodý: idiosynkraticky kódovaný, ne vždy vyhovující standardům a se stovkou existujících různých typů souborů, z nichž se každý může potenciálně pochlubit více verzemi. Tento již tolik rozmanitý obsah se navíc v průběhu času neustále mění. Při dlouhodobé ochráně tištěných dokumentů, se pracuje pouze s několika předem definovanými formáty, které jsou získány na základě přesných a rozpoznatelných parametrů.[34]

Webové archivy ale řeší ještě jiné než jen technické výzvy. Jedná se zejména o otázky zodpovědnosti a otázky autorského práva. Díky decentralizaci internetu není určena přímá zodpovědnost za dlouhodobé uložení webového obsahu a není tedy zcela jasné, která instituce by měla jakou část internetu archivovat. Globální povaha internetu také znamená, že odpovědnost za jeho ochranu nespadá úhledně do tradice národních kategorií[35] jako v případě národních knihoven a národní tištěné produkce.

---

[33] DAY, Michael. *Collecting and preserving the World Wide Web* [online]. 2003 [cit. 2015-10-12]. Dostupné z: http://www.jisc.ac.uk/uploaded_documents/archiving_feasibility.pdf

[34] LAZORCHAK, Butch. Web Archive Preservation Planning. In: *Library of Congress* [online]. 18. 8. 2011 [cit. 2015-10-12]. Dostupné z: http://blogs.loc.gov/digitalpreservation/2011/08/web-archive-preservation-planning/

[35] Tamtéž jako 32.

Situaci také zhoršuje autorské právo, které je mnohdy zastaralé a nepamatuje na webový obsah. To s sebou přináší hlavně problémy se zpřístupněním uživatelům. „*V České republice Autorský zákon (č. 121/2000 Sb.) sice umožňuje vytváření digitálního archivu, avšak jeho zpřístupňování nikoli.*“[36] To v praxi znamená, že v České republice může být celý webový archiv zpřístupněn pouze na půdě Národní knihovny na izolovaných počítačových stanicích.

[36] WebArchiv: získávání, archivace a zpřístupnění domácích webových zdrojů. *Ikaros* [online]. 2004, roč. 8, č. 5/2 [cit. 2015-10-12]. Dostupné z: http://www.ikaros.cz/node/1638

3.3. Metadata pro dlouhodobou ochranu

Metadata pro dlouhodobou ochranu jsou hlavní komponentou pro dlouhodobou digitální ochranu a jsou definována jako: nejrůznější typ dat, která umožňují opětovné vytvoření a interpretaci struktury a obsahu digitálních dat napříč časem.[37] Z této definice plyne, že existuje více typů metadat a podle organizace NISO[38] jsou to tři hlavní typy:[39]

- Popisná metadata jsou metadata, která popisují zdroj pro potřeby nalezení a identifikace. Nejčastěji obsahují informace o autorovi, titulu nebo předmětu.
- Strukturální metadata popisují digitální dokument z hlediska jeho struktury, tzn. jak digitální objekty složit dohromady, aby výsledkem byl zpět digitální dokument, např. posloupnost stran u zdigitalizované knihy.
- Administrativní metadata jsou nejširší oblastí metadat. Obsahují informace administrativního a technické charakteru, jako technický popis souborového formátu, informace o tom, kdo a kdy digitální dokument vytvořil atd. Také mohou obsahovat metadata, která se zabývají duševním vlastnictvím digitálního dokumentu. Ale zejména pod administrativní metadata patří metadata pro dlouhodobou ochranu.

Pro všechny tři typy metadat existují metadatové standardy, které slouží k nejrůznějším účelům a pro nejrůznější typy dokumentů. Pro digitální knihovny a archivy se nejčastěji využívají standardy napsané ve značkovacím jazyce XML. Příkladem takového standardu může být METS, který slouží pro *„zakódování deskriptivních, administrativních a strukturálních metadat popisující digitální objekty v repozitáři."*[40] Nicméně pro účely této práce jsou nejdůležitější metadata pro dlouhodobou ochranu, která slouží pro zaznamenání veškerých aktivit určených k zajištění dlouhodobé použitelnosti digitálního objektu.[41] Nejvýznamnějším metadatovým standardem pro tuto oblast je PREMIS, který je používán po celém světě napříč institucemi.

---

[37] DAY, Michael. The Long-Term Preservation of Web Content. MASANÈS, Julien. *Web archiving*. New York: Springer, c2006, s. 177-199. ISBN 3540233385-.

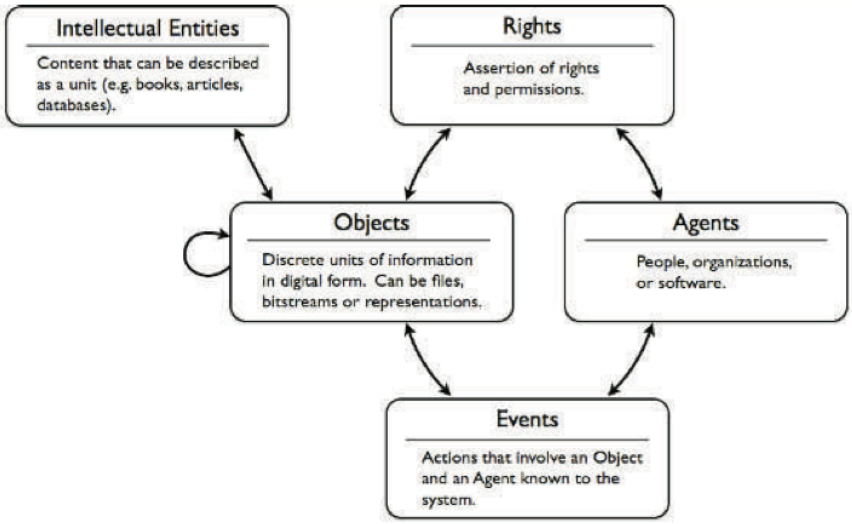
[38] NISO (National Information Standards Organization) je americká organizace spravující technické standardy

[39] ZENG, Marcia L. 3. Metadata types and functions. MARCIA L., Zeng. *Metadata Basics* [online]. 2007 [cit. 2015-10-12]. Dostupné z: http://marciazeng.slis.kent.edu/metadatabasics/types.htm

[40] CUBR, Ladislav. *Dlouhodobá ochrana digitálních dokumentů*. 1. vyd. Praha: Národní knihovna České republiky, 2010, 154 s. ISBN 978-80-7050-588-5.

[41] CAPLAN, Priscilla. *Understanding PREMIS* [online]. U.S.A.: The Library of Congress, 2009 [cit. 2015-10-12]. Dostupné z: http://www.loc.gov/standards/premis/understanding-premis.pdf

Metadatový standard PREMIS definuje základní sadu sémantických jednotek, které by měl repozitář znát, aby mohl plnit funkci dlouhodobé ochrany.[42] Mezi tyto funkce patří schopnost zajištění, aby digitální objekty bylo možné přečíst z médií a zobrazit je pomocí správného softwaru, a to vše při zachování jejich autenticity a bez nechtěných změn. Případné nutné změny musejí být zdokumentovány a zaznamenány, k čemuž také slouží PREMIS. PREMIS přináší vlastní datový model a definuje pět základních entit (obrázek č.1: datový model standardu PREMIS[43]), se kterými pak dále pracuje.



Obrázek č. 1: datový model standardu PREMIS

První entitou je intelektuální entita, která představuje intelektuální obsah jako jednotku, která je spravována (např. kniha), entita *Objects* pak představuje ten konkrétní digitální objekt, který je uložen v repozitáři. Tato entita pak může obsahovat informace jako unikátní identifikátor, informace o souborovém formátu, informace o tvorbě nebo například vztahy k dalším digitálním objektům.

Dalšími entitami jsou *Events* a *Agents*, které obsahují informace o událostech, které byly s objektem prováděny, a také, kdo je za tyto události zodpovědný. Například v případě digitalizace dokumentu může být jako událost uvedeno skenování a jako *Agents* by pak mohli být označeni skener a operátor skeneru. V případě webového obsahu se zde může uvést například crawler, který webový obsah sklidil.

---

[42] CAPLAN, Priscilla. *Understanding PREMIS* [online]. U.S.A.: The Library of Congress, 2009 [cit. 2015-10-12]. Dostupné z: http://www.loc.gov/standards/premis/understanding-premis.pdf

[43] Tamtéž jako 41

Rights je poslední entitou, která může obsahovat informace o omezeních pro práci s digitálním objektem, které plynou z autorského zákona. Jednou z informací, kterou může entita obsahovat, je např. čas, kdy vyprší duševní ochrana digitálního objektu, nebo jméno kolektivního správce, který za autorská práva odpovídá.

Konkrétní použití metadatového standardu PREMIS s příklady je uvedeno v poslední části práce, která se zabývá dlouhodobou ochranou webového obsahu v České republice.

3.4. Výčet strategií dlouhodobé ochrany

V oblasti dlouhodobé ochrany existuje celá řada strategií, které mají zabránit ztrátě dat. Bohužel pro webový obsah není možné aplikovat všechny, což je způsobeno výše zmiňovanou specifičností dat. Základní strategie dlouhodobé ochrany digitálních dat jsou:

- Migrace – „*Migrace ve smyslu dlouhodobé ochrany představuje konverzi digitálního objektu z originálního (zastaralého) formátu do formátu nového, který je podporován současnými technologiemi.*“[44] Podrobněji o migraci v kap. 3.5.
- Emulace – „*Emulací se rozumí zachování původního datového formátu digitálního objektu se znovuvytvořením některých procesů (např. hardwarové konfigurace nebo softwarových aplikací). Tímto způsobem je umožněno zobrazení dokumentu na současných počítačích.*“[45] Podrobněji o emulaci v kap. 3.6.
- Zapouzdření (Encapsulation) – „*Zapouzdření je technika, při níž se zapouzdří digitální objekt spolu se všemi nezbytnými prvky, které zajistí pozdější přístup k digitálnímu objektu.*“[46] Jedná se tedy o strategii, kdy k digitálnímu objektu je přidán kompletní popis jeho souborového formátu, ze kterého je možné i v budoucnosti rekonstruovat, jak souborový formát číst, např. jak znovu naprogramovat software, který souborový formát přečte. Při dlouhodobé ochraně webového obsahu je tato strategie nepoužitelná, neboť webový obsah je tvořen obrovským množstvím nejrůznějších souborových formátů. Navíc tuto strategii je možné aplikovat pouze na otevřené formáty, které nejsou patentově chráněny, a jejich plná specifikace je veřejně dostupná.
- Technologické muzeum – „*Pod pojmem technologické muzeum se rozumí deponování digitálního záznamu v podstatě jako artefakt v originálním formátu a prostředí a na originálním nosiči.*“[47] Tato technika není opět pro webový obsah použitelná, což je dáno jednak dynamickým vývojem webových technologií a jednak počtem těchto technologií, které jsou na internetu využívány. V případě webového obsahu by musel

---

[44] KRATOCHVÍLOVÁ, Zuzana. Dlouhodobá ochrana a zpřístupnění dat z webových archivů: WebArchiv Národní knihovny České republiky. *Knihovna: knihovnická revue*. 2012, roč. 23, č. 2., s. 35-47. Dostupné z:http://knihovna.nkp.cz/knihovna122/kratochv.htm

[45] LUKŠŮ, Alžběta. Dlouhodobé uchovávání digitálních dokumentů. In: *WikiKnihovna: Knihovníci sobě* [online]. 26. 4. 2010, 6. 2. 2012 [cit. 2015-10-12]. Dostupné z: http://wiki.knihovna.cz/index.php?title=Dlouhodobé_uchováván%C3%AD_digitáln%C3%ADch_dokumentů#Pou.C5.BEit.C3.A9_zdroje

[46] HLOUŠEK, Petr. *Problematika dlouhodobého uchovávání digitálních dat*. Brno, 2008. Dostupné z: http://is.muni.cz/th/179500/ff_b/BakalarskaDP.pdf. Bakalářská práce. Masarykova univerzita, Filozofická fakulta, Ústav české literatury a knihovnictví,, Kabinet informačních studií a knihovnictví.

[47] VOJTÁŠEK, Filip. Dlouhodobá archivace digitálních dokumentů. *Ikaros* [online]. 2000, roč. 4, č. 10 [cit. 2015-10-12]. Dostupné z: http://www.ikaros.cz/dlouhodoba-archivace-digitalnich-dokumentu

být uchován software jako internetové prohlížeče v různých verzích, všechny jejich možné plug-iny a dokonce i operační systém, na kterém půjde internetový prohlížeč spustit. To by pak, díky rychlému vývoji, vedlo k nutnosti uchovat příliš mnoho verzí hardwaru a softwaru s nejistým výsledkem. Proto tato strategie pro webový obsah představuje příliš velké riziko s neúměrnými náklady.

- Konverze do analogové formy (Hard Copying) – „*Strategie konverze digitálních dokumentů do analogové formy se opírá o skutečnost, že ochranné metody aplikované u analogových dokumentů jsou dostatečně ověřeny.*“[48] Tato strategie je zde uvedena pouze pro kompletní výčet strategií. Pro webový obsah vůbec nepřichází v úvahu. Jednou ze základních vlastností obsahu na internetu je jeho propojenost pomocí hypertextových odkazů a to samozřejmě na analogový nosič nelze přenést. Dalšími argumenty proti této strategii jsou nemožnost konvertovat videa, animace, ale i interaktivita, kterou zkušenost používání internetu s sebou přináší.

Všechny strategie z výčtu je možné mezi sebou různě kombinovat, ať už horizontálně (např. pro část dat využít migraci a pro další část emulaci) nebo vertikálně (nejprve může být zvoleno zapouzdření a za nějaký čas na základě zapouzdřených informací vytvořit migrační nástroje). V současné době jsou strategie použitelné pro webový obsah pouze migrace a emulace, případně jejich kombinace.

[48] VOJTÁŠEK, Filip. Dlouhodobá archivace digitálních dokumentů. *Ikaros* [online]. 2000, roč. 4, č. 10 [cit. 2015-10-12]. Dostupné z: http://www.ikaros.cz/dlouhodoba-archivace-digitalnich-dokumentu

3.5. Migrace

Strategie, která se ukazuje jako jedna z možných pro dlouhodobou ochranu webového obsahu, je migrace. Jak bylo napsáno výše ve výčtu strategií dlouhodobé ochrany, migrace představuje převod digitálního objektu do jiného formátu, u kterého se předpokládá, že je modernější, a tudíž má větší technologickou podporu a nehrozí u něj v blízké době zastarání. Účelem strategie migrace je kontinuálně převádět digitální objekty na novou generaci softwaru a hardwaru.[49] Předpokladem pro tuto činnost je, že s novější (současnou) generací softwaru a hardwaru je jednodušší pracovat, neboť je stále podporována, vyvíjena a používána.

Aktuálnost souborového formátu není jediný požadavek, který musí nový formát splnit. „*Úspěšná migrace vyžaduje detailní analýzu a identifikaci významných vlastností objektů, které by měly zůstat zachovány i při přechodu na nový formát.*“[50] Teoreticky by tedy bylo možné webový obsah migrovat například do formátu PDF, kdy by zůstala zachována struktura stránek a obsah, ale také by se tím ztratila určitá uživatelská zkušenost, která procházení webových stránek doprovází.

Možná ztráta významných vlastností je uváděna jako jedna z velkých nevýhod migrační strategie. V některých případech je velmi těžké zvolit správný formát, do kterého migrovat. S tím také souvisí postup migrace, který naráží na komplexnost webových stránek. Není možné migrovat webovou stránku jako celek, ale je nutné migrovat jednotlivé objekty, ze kterých se webová stránka skládá.

Tato skutečnost přináší další problém, a tím je zachování odkazů na objekty. Pokud je na webové stránce umístěn například obrázek ve formátu JPEG a ten bude migrován do jiného formátu, pak webová stránka bude i nadále odkazovat na starý obrázek ve formátu JPEG. Tento problém řeší kontejnerový formát WARC, který umožňuje v metadatech referovat na nové migrované soubory a pomocí těchto referencí se pak řídí zobrazovací aplikace.

Jaký je tedy správný postup migrace? Postup migrace webového obsahu lze rozdělit do čtyř základních kroků. Tento postup se týká pouze kontejnerového formátu WARC,

---

[49] DAY, Michael. The Long-Term Preservation of Web Content. MASANÈS, Julien. *Web archiving*. New York: Springer, c2006, s. 177-199. ISBN 3540233385-.

[50] KVAŠOVÁ, Zuzana a Tomáš SVOBODA. Dlouhodobá ochrana elektronických publikací. *ProInflow* [online]. 2013, Vol. 5, No. 2 [cit. 2015-03-03]. Dostupné z: http://www.phil.muni.cz/journals/index.php/proinflow/article/view/775

který s migrací počítá. Provádět migraci souboru uvnitř formátu ARC není příliš vhodné, neboť starší formát si s novými soubory neporadí a bylo by nutné měnit zdrojové kódy webových stránek. Proto je výhodnější nejprve provést migraci kontejnerových formátů.[51]

Čtyři kroky migrace:[52]

1. Preservation Planning – Před započetím samotné migrace je potřeba nejprve identifikovat a vyhodnotit akce dlouhodobé ochrany. To znamená zvolit vhodný nový souborový formát a nástroj migrace, vyhodnotit potenciální rizika a případné ztráty vlastností oproti původnímu formátu.

2. Identification, Extraction and Validation – Ve druhém kroku přichází na řadu identifikace objektů v kontejnerech a jejich extrakce. S tím souvisí i validace extrahovaných souborů, tedy ujištění se, že jde o formát, který má být migrován, neboť pouze přípona souboru toto nemusí garantovat.

3. Preservation Action – Ve třetím kroku přichází na řadu samotná migrace, kdy za použití specializovaného nástroje dojde k transformaci z jednoho formátu do jiného. Po transformaci musí opět proběhnout validace, že nový soubor není poškozen.

4. Injection – Posledním krokem je vložení nových souborů do nového WARC kontejneru. Nový soubor se musí zapsat včetně nových metadat a referencí na původní soubor. U původního souboru se pak jen v metadatech také zapíše reference na nový objekt. Po zapsání všech nových souborů se provede validace kontejneru jako celku.

Reference mezi jednotlivými verzemi souborů jsou realizovány pomocí hlaviček, které v kontejneru WARC obsahuje každý digitální objekt. V principu se jedná o to, že pokud si uživatel vyžádá po zobrazovací aplikaci zobrazení digitálního objektu, tak zobrazovací aplikace sáhne do kontejneru, kde objekt leží. V případě žádosti o zastaralý objekt se aplikace z metadatové hlavičky dozví, že tento objekt je zastaralý a dostane

---

[51] Migrace mezi formáty ARC a WARC ovšem také není bez rizika a přestože na ni existují nástroje, musí každý archiv zvážit veškerá pro i proti.

[52] STRODL, Stephan, Peter Paul BERAN a Andreas RAUBER. Migrating Content in WARC Files. In: MASANES, Julien a Andreas RAUBER. *The 9th International Web Archiving Workshop (IWAW 2009) Proceedings* [online]. Paris (France): European Archive Foundation, 2009, s. 43-49 [cit. 2015-03-03]. Dostupné z: http://publik.tuwien.ac.at/files/PubDat_181115.pdf

identifikátor nástupce. Pomocí identifikátoru si pak najde, kde nový objekt leží, a ten pak zobrazí uživateli.

Tento systém referencí byl vymyšlen při tvorbě WARC formátu a je ukotven v jeho ISO standardu[53] ve formě speciálního záznamu s názvem *conversion*, který je typicky využíván k uložení obsahu z migrace, která byla provedena za účelem udržení životaschopnosti obsahu, např. pokud běžně dostupné zobrazovací nástroje pro původní souborový formát přestanou být standardním vybavením koncového uživatele. Podle potřeby může být původní obsah migrován do životaschopnější formátu, s cílem udržet informace použitelné se současnými nástroji.[54]

Z uvedeného postupu migrace je zřejmé, že celý proces s sebou nese určitá rizika a samozřejmě nároky na počítačovou infrastrukturu a programové vybavení. Není potřeba jen nástroj pro samotnou migraci, ale také nástroj pro validaci a charakterizaci formátu, a to i pro finální validaci kontejnerového formátu. To s sebou nese nároky na výpočetní výkon s ohledem na obrovské objemy dat webových archivů. A při tak velkém počtu operací je tu velké riziko chybovosti.

[53] Standard pro formát WARC se jmenuje ISO 28500:2009 – WARC file format.

[54] ISO 28500:2009. *WARC file format*. 1. vyd. Londýn: British Standard Institute, 2009

3.6. Emulace

Druhou strategií, kterou je možné využít pro dlouhodobou ochranu webového obsahu je emulace. Emulace se na rozdíl od migrace nezaměřuje na samotný digitální objekt, ale na technologické prostředí, ve kterém byl objekt vytvořen.[55] Emulaci lze definovat jako napodobení technologického prostředí nebo jeho části, ze kterého pochází digitální objekt, pomocí speciálních nástrojů za účelem zobrazení tohoto objektu v současném technologickém prostředí.

To znamená, že pomocí speciálních nástrojů je simulován starý software nebo i hardware tak, aby byl zastaralý digitální objekt zobrazený v prostředí, které vypadá jako jeho originální. Je tedy zřejmé, že nedochází k žádné modifikaci samotného objektu, jako je tomu v případě migrace. Samotná emulace může být prováděna na třech základních úrovních:[56]

1. Na úrovni aplikace: vytvoření aplikace, která dělá přesně to samé jako aplikace původní. Například vytvoření napodobeniny textového editoru pro zastaralý formát T602, který bude spustitelný na nejnovějších operačních systémech.

2. Na úrovni operačního systému: vytvoření programu, který simuluje zastaralý operační systém, ve kterém pak bude možné spouštět původní aplikace. Například emulátor operačního systému DOS, ve kterém bude spuštěn původní textový editor Text602, který pracuje s formátem T602.

3. Na úrovni hardwarové architektury: vytvoření programu, který slouží jako platforma, na které bude moci být spuštěn původní operační systém s původními aplikacemi. Příkladem může být program, který simuluje starou architekturu, na které může být nainstalovaný operační systém DOS s textovým editorem Text602 pro zobrazení formátu T602. Tento přístup je nejčastější, neboť není nutné emulovat aplikace nebo operační systémy, ale stačí zachovat ty původní.[57]

Při budování emulátoru pro potřeby dlouhodobé ochrany je nutné vzít v potaz dva aspekty přístupu k emulaci, těmi jsou modularita a trvanlivost. Modulační strategie

---

[55] Selecting the right preservation strategy. *Paradigm: The Personal Archives Accessible in Digital Media* [online]. 2008 [cit. 2015-03-03]. Dostupné z: http://www.paradigm.ac.uk/workbook/preservation-strategies/selecting-emulation.html

[56] ISO 28500:2009. *WARC file format*. 1. vyd. Londýn: British Standard Institute, 2009

[57] Tamtéž jako 54.

emulace je definována jako systém jednotlivých menších emulátorů, které emulují jednotlivé části technologického prostředí, jsou vzájemně propojeny, a tím vytváří celkový emulační proces.[58] V praxi to vypadá tak, že jsou emulovány jednotlivé komponenty hardwaru zvlášť, a pak jsou sestaveny dohromady, obdobně jako fyzický hardware. Stejně jako u fyzického hardwaru je také možné komponenty, resp. emulátory různě obměňovat.

Přístup budování emulátoru zaměřený na trvanlivost (ang. durability) vychází z premisy, že každá aplikace je závislá na nižší platformě skládající se z hardwaru a softwaru.[59] To znamená, že změna platformy, na které běží aplikace, může mít negativní dopad na funkčnost aplikace. Vytvořit aplikaci, která by fungovala navždy napříč platformami, není možné, protože každá aplikace je na své platformě závislá. Při budování emulátoru zaměřeného na trvanlivost je vytvořena tzv. mezivrstva, která je vložena mezi aplikaci a její platformu. Nazývá se virtual machine (dále jen VM). VM slouží k podpoře aplikace, aby mohla fungovat na různých platformách bez nutnosti změny aplikace.[60]

Aplikace komunikuje jen s VM, které se vydává za podporovanou platformu a zajišťuje komunikaci s novým hardwarem a softwarem. Do budoucna stačí upravovat pouze VM, aby se uměl spojit s další novou platformou. Tento přístup je velmi výhodný, když více aplikací, které je potřeba spustit, jsou ze stejné platformy. V takovém případě stačí pouze jedna VM.

Dále již k samotným výhodám a nevýhodám emulace. Největší výhodou emulace je, že nedochází k žádné modifikaci původního souboru, ale ani k žádné jeho manipulaci. Není tu žádné riziko ztráty[61]. V případě velkých kolekcí (jako mají právě webové

---

[58] HOEVEN, Jeffrey Van der, Bram LOHMAN a Remco VERDEGEM. Emulation for Digital Preservation in Practice: The Results. *International Journal of Digital Curation* [online]. Bath: UKOLN, University of Bath, 2007, vol. 2, issue 2, s. 207-219 [cit. 2015-03-03]. DOI: 10.1007/978-3-540-33640-2_10. Dostupné z: http://50.17.193.184/omeka/files/original/84cca606bbb8f1955f42b22c29268811.pdf

[59] Selecting the right preservation strategy. *Paradigm: The Personal Archives Accessible in Digital Media* [online]. 2008 [cit. 2015-03-03]. Dostupné z: http://www.paradigm.ac.uk/workbook/preservation-strategies/selecting-emulation.html

[60] Tamtéž jako 59.

[61] LONG, Andrew Stawowczyk. *Long-term preservation of web archives: Experimenting with emulation and migration methodologies* [online]. Australia: National Library of Australia, 2009 [cit. 2015-03-03]. Dostupné z: http://www.netpreserve.org/sites/default/files/resources/Methodologies.pdf

archivy) se emulace jeví jako mnohem levnější řešení než migrace.[62] A to hlavně díky tomu, že emulace může být provedena už na hardwarové úrovni, teoreticky by bylo možné vytvořit jeden emulátor pro celou kolekci. Emulace je také jediné řešení pro specializované, obskurní a velmi složité formáty, které jsou při migraci logicky ignorovány. Emulace také v duchu myšlenky dlouhodobé archivace přináší původní uživatelský zážitek při prohlížení objektu.

Velkou nevýhodou emulace je, že znemožňuje nebo dělá velmi složitou výměnu dat mezi systémy. Je složité dostat informace z emulátoru do současného systému v takové formě, aby se s nimi dalo dále pracovat. Vyvinout komplexní emulátor je podstatně náročnější než migrační nástroj. Emulace je také několikanásobně dražší než migrace.[63] Při budování emulátoru je nutností detailně znát emulované prostředí[64], a to může být zpětně velmi složité i díky tomu, že některé komponenty nemusí mít volně dostupnou dokumentaci nebo nemusí mít vůbec žádnou. Většina emulačních strategií představuje nákup a uchování proprietárních aplikací, kterými budou formáty v emulátoru zobrazovány.[65] Tyto aplikace pak mohou mít časově omezenou licenci, mohou vyžadovat aktivaci u výrobce (ten již nemusí existovat) nebo mohou obsahovat jinou protipirátskou ochranu. Zásadním problémem je také to, že vytvořený emulátor, je také jen aplikace, která časem zastará.

Tím se dostáváme k dalšímu problému, a tím je emulace v čase. Pokud je jednou zvolena emulace jako hlavní strategie dlouhodobé archivace, pak nastává otázka, jakým způsobem zajistit, že emulátor bude fungovat na dalších generacích technologií.[66] Jednou z možností vyřešení problému je použití VM, a pouze její udržování. Druhá možnost je přepisování celého emulátoru, tato je ale o poznání nákladnější. Třetí možností je řetězení

---

[62] Selecting the right preservation strategy. *Paradigm: The Personal Archives Accessible in Digital Media* [online]. 2008 [cit. 2015-03-03]. Dostupné z: http://www.paradigm.ac.uk/workbook/preservation-strategies/selecting-emulation.html

[63] ROSENTHAL, David S.H. Emulation & Virtualization as Preservation Strategies. In: The Andrew W. Mellon Foundation [online]. New York: The Andrew W. Mellon Foundation, 2015 [cit. 2016-07-26]. Dostupné z: https://mellon.org/Rosenthal-Emulation-2015

[64] LONG, Andrew Stawowczyk. *Long-term preservation of web archives: Experimenting with emulation and migration methodologies* [online]. Australia: National Library of Australia, 2009 [cit. 2015-03-03]. Dostupné z: http://www.netpreserve.org/sites/default/files/resources/Methodologies.pdf

[65] Tamtéž jako 63.

[66] DAY, Michael. The Long-Term Preservation of Web Content. MASANÈS, Julien. *Web archiving*. New York: Springer, c2006, s. 177-199. ISBN 3540233385-.

emulátorů (ang. chaining), zjednodušeně jde o metodu, kdy se pro emulátor vytvoří jiný emulátor, umožňující předchozí emulátor spustit a tak dále.

Emulace v čase je o to složitější, čím starší je webový archiv. Otázka, která se nabízí: kolik by bylo potřeba emulovat verzí operačních systémů, resp. verzí internetových prohlížečů, když v dnešní době je běžné, že má webový archiv data stará deset a více let? Za posledních deset let tu bylo hned několik dominantních internetových prohlížečů v různých verzích a to včetně plug-inů jako je Flash player, Silverlight apod.

Současným trendem na poli emulace je tvorba emulačních frameworků[67], které běží na serverech dostupných z internetu (na cloudu), a k těmto frameworkům je možné přistupovat přes webový prohlížeč, případně přes API[68], jak je běžné u jiných internetových služeb. Emulátor jako internetová služba přináší: zajištění emulace na vyžádání (ang. on-demand), jednotné rozhraní pro různé emulátory, unifikovaný síťový přístup.[69] Tyto vlastnosti emulátoru přináší jednodušší použitelnost pro uživatele bez nutnosti instalace nebo tvorby vlastního emulátoru a mohou přispět ke snadnějšímu zapojení emulace ve strategiích dlouhodobé ochrany.

Přímo pro webové archivy vznikl specializovaný framework zvaný oldweb.today, což je odlehčený emulační framework webových prohlížečů, který slouží k propojení webových archivů se staršími verzemi webových prohlížečů. Oldweb.today neemuluje přímo hardware, ale je to kombinace reverzního inženýrství a volně dostupných zdrojových kódů, které umožňují spuštění starých prohlížečů na systému Linux.[70] Oldweb.today je i webová služba, které je veřejně dostupná pro běžného uživatele, který si tak může vyzkoušet, jak vypadaly webové prohlížeče a v nich webové stránky v minulosti.

---

[67] Framework (aplikační rámec) je softwarová struktura, která slouží jako podpora při programování a vývoji a organizaci jiných softwarových projektů.

[68] API (Application Programming Interface) označuje v informatice rozhraní pro programování aplikací. Jde o sbírku procedur, funkcí, tříd či protokolů nějaké knihovny, které může programátor využívat. API určuje, jakým způsobem jsou funkce knihovny volány ze zdrojového kódu programu.

[69] Emulation as a Tool for Web Preservation: Authentic Access and Efficient Web-server Preservation. International *Internet Preservation Consortium* [online]. 2016 [cit. 2016-07-19]. Dostupné z: http://www.netpreserve.org/sites/default/files/WAC-Thomas_Liebetraut.pdf

[70] Cyberspace, the old-fashioned way. *Rhizome* [online]. NY, USA: New Museum, 2015 [cit. 2016-07-19]. Dostupné z: http://rhizome.org/editorial/2015/nov/30/oldweb-today/

3.7. Shrnutí

Pro migraci i emulaci platí nutnost dokonalé znalosti archivu, na který budou strategie dlouhodobé ochrany aplikovány. U migrace se jedná o znalost souborového formátu objektů a u emulace pak prostředí, ve kterém digitální objekty vznikaly. Pro obě platí, že ani jeden přístup není natolik univerzální, aby dokázal být jedinou strategií pro ochranu celého archivu. Vždy je nutné vyhodnotit všechna rizika a náklady, které s sebou aplikace strategie nese.

Velmi také záleží na účelu archivu, tzn. co archiv chce nabízet svým uživatelům a hlavně kdo jsou jeho uživatelé a co od archivu očekávají. V případě univerzálních webových archivů, které se snaží uchovat co nejkomplexnější část internetu pro své uživatele, jako tomu bývá u národních webových archivů, nezbývá nic jiného, než správně zvolit kombinaci obou strategií dlouhodobé ochrany. Klíčové při rozhodování, kterou strategii nebo kombinaci strategií zvolit, je pochopit, že účelem každé strategie je zajistit, aby všechny signifikantní vlastnosti objektu zůstaly uchovány.[71] Díky vzniku emulačních frameworků, které velmi ulehčují práci s emulátory, je dnes emulační strategie trendem, na který se zaměřují experti z nejrůznějších webových archivů. Nicméně je třeba na závěr této kapitoly zmínit, že veškeré snahy o zapojení emulace do procesů dlouhodobé ochrany jsou zatím jen ve fázi experimentů.

[71] DAY, Michael. The Long-Term Preservation of Web Content. MASANÈS, Julien. *Web archiving*. New York: Springer, c2006, s. 177-199. ISBN 3540233385-.

4. Současné nástroje pro dlouhodobé uchování webového obsahu

V této části práce je zmíněno pouze několik institucí nebo organizací, které se věnují dlouhodobé ochraně webového obsahu. Výčet institucí není kompletním výčtem ani reprezentativním vzorkem. Vybrány byly instituce, se kterými spolupracoval v minulosti autor práce nebo webový archiv NK ČR. Tyto instituce se vyznačují tím, že přinášejí inovace na poli dlouhodobé ochrany. Stručný výčet je pouze jakýmsi vrcholem ledovce, který má naznačit, jakým směrem se ve světě ubírá vývoj v oblasti dlouhodobé ochrany. V dnešní době se začalo dlouhodobou ochranou webového obsahu zabývat mnoho institucí, je to logický krok pro všechny archivy v případě, že mají již vyřešenou jeho akvizici.

Výběr institucí má pomoci ilustrovat tvrzení uvedené v teoretické části, ale zároveň má také sloužit k ukázce toho, že jeden problém má mnoho řešení a že zároveň v současné době neexistuje žádné univerzální. Ve výčtu je jedno celosvětové konsorcium, ve kterém se webové archivy snaží organizovat a společně postupovat při tvorbě strategií dlouhodobé ochrany. Také jsou zde zastoupeny dvě evropské knihovny, které vyvíjejí své vlastní nástroje a jsou leadery inovací v oblasti dlouhodobé ochrany webového obsahu.

Poslední kapitola této části je věnovaná problematice webu 2.0 a jeho dlouhodobé ochraně. S institucí se této problematice věnuje Kongresová knihovna ve Washingtonu s projektem Twitter archivu. Nicméně u archivace webu 2.0, není ještě úplně vyřešena jeho akvizice, a tak samotná dlouhodobá ochrana je teprve na začátku.

4.1. International Internet Preservation Consortium (IIPC)

International Internet Preservation Consortium (IIPC) je mezinárodní organizace sdružující převážně webové archivy, případně instituce webové archivy provozující. Konsorcium bylo založeno již v roce 2003 ve Francii a český webový archiv v zastoupení Národní knihovny ČR je členem od roku 2007. Cílem sdružení je spolupráce na tvorbě nástrojů, standardů a vytváření nejlepší praxe v oblasti webového archivnictví.[72]

Nejjednodušeji lze shrnout strategii IIPC pro dlouhodobou ochranu webového obsahu jako spolupráci. IIPC vychází z teze, že web je mezinárodní, a tak dává největší smysl pracovat na jeho ochraně na mezinárodní úrovni.[73] K tomuto účelu si konsorcium stanovilo tři cíle, které mají k dlouhodobé ochraně webu přispět:[74]

1. Umožnit akvizici velké části webového obsahu z celého světa tak, aby byla zajištěna jeho dlouhodobá ochrana včetně zpřístupnění.

2. Podporovat rozvoj a využívání společných nástrojů, metodik a standardů, které umožní vytváření mezinárodních archivů.

3. Povzbuzovat a podporovat národní knihovny po celém světě ve věci archivace webového obsahu.

O plnění těchto cílů se snaží IIPC již od svého založení a po tuto dobu se mu to daří s větším nebo menším úspěchem. V současné době má IIPC k dispozici vývojáře, kteří rozvíjí nástroje, dále několik mezinárodních odborných pracovních skupin, které pracují na vytváření metodik, standardů a hledání řešení pro objevující se problémy. V neposlední řadě se IIPC zabývá výzkumem národních knihoven, na základě kterého připravuje nejrůznější školení, workshopy a další edukační aktivity.

IIPC si například vzala pod svou správu open-source verzi aplikace Wayback Machine, která se nazývá OpenWayback, Její vývoj konsorcium zaměřilo více na potřeby a problémy svých členů, na rozdíl od Internet Archive, který vyvíjí Wayback Machine

---

[72] About IIPC. *IIPC: International Internet Preservation Consortium* [online]. 2012 [cit. 2015-03-03]. Dostupné z: http://netpreserve.org/about-us

[73] GOETHALS, Andrea, Clément OURY, David PEARSON, Barbara SIERMAN a Tobias STEINKE. Facing the Challenge of Web Archives Preservation Collaboratively: The Role and Work of the IIPC Preservation Working Group. D-Lib Magazine [online]. 2015, **21**(5/6), - [cit. 2016-07-26]. DOI: 10.1045/may2015-goethals. ISSN 1082-9873. Dostupné z: http://www.dlib.org/dlib/may15/goethals/05goethals.html

[74] Tamtéž jako 72.

pro svoje využití. IIPC se zároveň snaží implementovat veškeré úpravy, které provádí Internet Archive, aby se obě verze aplikace od sebe příliš nelišily.

Dalším velkým projektem IIPC je agregátor světových archivů zvaný Memento, který uživatele naviguje do webového archivu, jenž má uživatelem požadovaný zdroj k dispozici.

IIPC působí v celé šíři oblasti webové archivace, tedy i v oblasti dlouhodobé ochrany webových zdrojů, pro kterou má vytvořenou odbornou pracovní skupinu s názvem Preservation Working Group, která se snaží přinést nové poznatky, doporučení, standardy a výzkum právě v oblasti dlouhodobé archivace.

V rámci konsorcia IIPC je zcela běžné, že se některé členské instituce domluví mezi sebou a vyvíjejí nebo se podílejí na vývoji nástrojů pro práci s archivy. Ne vždy musí být nutně zapojeny všichni nebo většina členů. Přestože mají webové archivy stejný hlavní cíl, jejich potřeby mohou být zcela odlišné. Velký rozdíl plyne z autorskoprávní legislativy: anglosaská x evropská legislativa; Legal deposit[75]; implementace knihovního zákona a další. Rovněž záleží na zaměření archivu – jiné potřeby pro práci s daty bude mít archiv národní knihovny a jiné malý univerzitní archiv. Každý webový archiv má nastavené své priority, do kterých chce aktuálně investovat své, často velmi omezené, finanční nebo i personální prostředky.

Velmi často dochází k situaci, kdy nějaký nástroj vyvine několik členů, kteří aktuálně řeší nějaký problém a v budoucnosti jej převezmou další členské instituce, když se setkají s podobným problémem. Tyto menší nástroje a aplikace bývají nezřídka implementovány do systémů pro dlouhodobou ochranu dat. Jedná se například o migrační nástroje WARC Tools nebo HTTrack2ARC, které slouží k transformaci dat mezi kontejnerovými formáty, dále validační nástroje pro kontejnerové formáty JWAT nebo JHOVE2. A také nástroje pro extrakci nebo prozkoumávání metadat, jako je WAT nebo WarcManager.

[75] Obdoba povinného výtisku pro digitální dokumenty.

4.2. Britská národní knihovna – Interject

Britská národní knihovna začala s archivováním webového obsahu v roce 2004 a v současné době je jedním z leaderů rozvoje v oblasti dlouhodobé archivace. Nastavení jejich archivu je zcela odlišné od toho českého. Pro dlouhodobou ochranu webového obsahu využívá pouze kontejnerový formát WARC tzn. starší data, která jsou uložená ve formátu ARC, při ukládání do systému dlouhodobé ochrany migrují do novějšího formátu.

Pro dlouhodobou ochranu si definovala tři různé intelektuálními entity[76], které se vztahují k webovým stránkám. K tomuto kroku mohla přistoupit, protože sklízí vždy jednu webovou stránku do jednoho kontejneru.[77] Takové sklízení lze označit za sériové, neboť každý jednotlivý sklízecí stroj má otevřený svůj vlastní kontejner, do kterého ukládá počítačové soubory stažené pouze z jedné webové stránky a až v okamžiku, kdy tuto činnost dokončí, přejde na novou stránku a pro ni otevře nový kontejner.

Britská národní knihovna definovala intelektuální entity: *website* (kolekce stránek, které patří pod stejnou doménu); *webpage* (jedna stránka, která je zobrazitelná jako objekt a vede na ni jeden nebo více hypertextových odkazů) a *associated objects* (je částí stránky, např. obrázek na stránce apod.)[78].

Nástroj, který Britská národní knihovna vytvořila pro implementaci strategií dlouhodobé ochrany se jmenuje Interject. Při jeho vytváření autoři vycházeli z teze, že v současné době se většina systémů dlouhodobé ochrany stará o ekonomické a technické stránky[79] a problém se zastaráváním obsahu je odsunut do pozadí. Zastaráváním je myšleno nemožnost otevřít soubor současným počítačovým vybavením. Velkým problémem je pak poznat, kdy se formát stal již zastaralým, zejména u specializovaných souborových formátů (např. formáty pro 3D grafiku).

---

[76] Více o intelektuální entitě kap. 5.2.

[77] ENDERS, Markus. A METS based information package for long term preservation of web archives. In: RAUBER., Andreas a Andreas RAUBER. *IPRES 2010 proceedings of the 7th International Conference on Preservation of Digital Objects ; September 19 - 24, 2010, Vienna, Austria*. Wien: Österreich. Computer Gesellsch, 2010, s. 31-40. ISBN 9783854032625. Dostupné z: http://publik.tuwien.ac.at/files/PubDat_191968.pdf

[78] Tamtéž jako 76.

[79] JACKSON, Andy. User driven digital preservation with Interject. THE BRITISH LIBRARY. *UK Web Archive blog* [online]. 2014 [cit. 2015-03-03]. Dostupné z: http://britishlibrary.typepad.co.uk/webarchive/2014/08/user-driven-digital-preservation-with-interject.html

Interject byl navržen tak, aby pomohl zastaralé formáty odhalit. Pracuje se dvěma teoriemi. Tou první je, že formát, který je plně otevřený a je známá jeho specifikace, která přesně popisuje, jak s formátem pracovat, nemůže být nikdy zastaralý. Druhá teorie pak říká, že pokud nějaký člen z určené skupiny uživatelů archivu nedokáže formát otevřít, tak je zastaralý.[80] To znamená, že plně otevřený formát by bylo možné díky tomu, že je známá jeho specifikace, otevřít i za dvacet let, ale už to neznamená, že jej za dvacet let nebo dokonce v tuto chvíli dokáže otevřít uživatel archivu.

Interject funguje na principu zpětné vazby. Pokud uživatel prohlížející si archiv narazí na soubor, který nedokáže otevřít, tak to jednoduše nahlásí. Interject vyhodnotí, jestli jde o již známý problém, v takovém případě nabídne řešení např. emulátor v internetovém prohlížeči. V případě že se jedná o zcela nový problém, předá Interject požadavek správci archivu.

Toto řešení slouží zejména k určení priorit v zavádění strategií dlouhodobé ochrany. Není to ale řešení, které by zcela nahradilo badatelskou nebo kurátorskou činnost, nicméně může být velmi užitečným nástrojem při rozhodování.

[80] JACKSON, Andy. User driven digital preservation with Interject. THE BRITISH LIBRARY. *UK Web Archive blog* [online]. 2014 [cit. 2015-03-03]. Dostupné z: http://britishlibrary.typepad.co.uk/webarchive/2014/08/user-driven-digital-preservation-with-interject.html

4.3. Francouzská národní knihovna – SPAR a ontologie

Francouzská národní knihovna (dále jen BNF) je jedním ze zakladatelů konsorcia IIPC a v oblasti webové archivace a dlouhodobé ochrany má velké zkušenosti. Na její půdě se vyvíjí modulární systém pro dlouhodobou ochranu SPAR (Scalable Preservation and Archiving Repository), který je založen na OAIS standardu, ale také přináší inovační procesy pro práci s daty.

Pro svá archivní data BNF vytvořila ontologický datový model, který je implementován pomocí běžných metadatových standardů (METS, PREMIS) a technologie RDF, ale zároveň je na těchto technologiích a standardech zcela nezávislý. Ontologický datový model pro SPAR je vytvořen tak, aby byl samopopisný. „*BNF navrhla vlastní implementaci formátu RDF, založenou primárně na převzetí některých částí několika zavedených ontologií (např. DC) pro popis prvků a jejich vztahů. Každý termín je jednoznačně identifikován schématem INFO:URI.*“[81] Při tvorbě ontologického datového modelu se BNF striktně držela referenčního modelu OAIS, a tak vytvořila pro každou kategorii metadat definovanou standardem OAIS vlastní ontologii.[82]

BNF tento model uložení dat primárně nevyvíjela pro data pocházející z webového archivu, ale ukázalo se, že velmi dobře funguje i pro takto specifický druh dat. Díky tomuto datovému modelu má BNF i pro tak velké objemy nestrukturovaných dat podobné možnosti správy jako u standardizovaných digitalizovaných dat. SPAR pak poskytuje BNF statistiky, ukazatele a prolinkování dat, které jsou nutné k jejich evaluaci.[83] Znalost dat v archivu je nezbytnou premisou pro zavádění akcí dlouhodobé ochrany.

BNF využila svůj stávající systém dlouhodobé ochrany, který byl původně vytvořen pro digitalizovaná data i pro data z webového archivu. Výhody, které tento přístup přináší, jsou zejména nižší finanční nároky, možnost využití již dříve vytvořených metodik a postupů a v neposlední řádě je jeden systém výhodnější na řízení a udržování.

---

[81] CUBR, Ladislav. *Zpráva ze služební cesty: Návštěva Národní knihovny Francie (BNF) [online]*. Praha, 2012 [cit. 2015-03-03]. Dostupné z: https://www.nkp.cz/soubory/ostatni/cz_pariz2012_lc.pdf

[82] BERMÈS, Emmanuelle a Gautier POUPEAU. Semantic Web technologies for digital preservation: the SPAR project. In: *Proceedings of the Poster and Demonstration Session at the 7th International Semantic Web Conference* [online]. 2008 [cit. 2015-03-03]. Dostupné z: http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-401/iswc2008pd_submission_14.pdf

[83] Preservation Is Knowledge: A community-driven preservation approach. In: *IPRESS 2012: Proceedings of the 9th International Conference on Preservation of Digital Objects* [online]. Toronto, 2012 [cit. 2015-03-03]. Dostupné z: http://www.bnf.fr/documents/ipress2012_art_spar.pdf

Nevýhodou pak může být nutnost využití již existujícího datového modelu, který je možné upravovat a vylepšovat, ale nemůže být nahrazen novým. [84]

[84] OURY, Clément a Sébastien PEYRARD. From the World Wide Web to digital library stacks: preserving the French web archives. In: Proceedings of the 8th International Conference on Preservation of Digital Objects (iPRES) [online]. Singapore: National Library Board : Nanyang Technological University, 2011, s. 237-241 [cit. 2016-07-26]. ISBN 978-981-07-0441-4. Dostupné z: https://halshs.archives-ouvertes.fr/halshs-00868729/document

4.4. Web 2.0 a Twitter archiv Kongresové knihovny

K pojmu web 2.0 lze najít několik definic, nicméně pro účely této práce postačí zjednodušení, že se jedná o web, jehož obsah je dynamicky tvořen, editován a mazán jeho uživateli. Problémy, které z toho plynou při snaze o jeho archivaci, se týkají hned několika oblastí: jednak oblasti technické, oblasti duševního vlastnictví a v neposlední řadě souvisejí i se specifiky z hlediska dynamičnosti a struktury obsahu.

Technické problémy představují technologie, které jsou využívány k vytváření funkčních a interaktivních částí, které web 2.0 potřebuje ke svému chodu (zejména dynamické načítání obsahu). Často se jedná o aplikace postavené na technologii AJAX[85], které využívají k změně obsahu asynchronní zpracování webových stránek. S těmi si neumí poradit sklízeče nebo aplikace pro zpřístupnění. Problém je taky v samotném interpretování ostatních složitějších JavaScriptových skriptů, neboť zapojení jejich interpretu[86] do procesu sklízení webových stránek výrazně navyšuje jeho časovou náročnost.

Další oblastí, do které web 2.0 přináší nové otázky k vyřešení je vlastnictví a zodpovědnost za obsah.[87] Pokud legislativa dané země omezuje webovému archivu provozovat některou z jeho činností (nejčastěji se jedná o zpřístupnění archivovaných zdrojů), pak je třeba žádat vlastníka o udělení práv, umožňujících se zdroji dále nakládat. U webových serverů, kde obsah nahrávají uživatelé, je pak těžké určit, na koho se s žádostí obrátit – komu náleží autorská práva za publikovaný obsah.

Hlavní problém ale přináší samotný charakter obsahu, který neustále přibývá a také v mnoha případech mizí, je uživateli postupně editován nebo má nastavený přístup pro různé skupiny uživatelů a ne vždy je zcela veřejný nebo je dostupný po registraci.

Největší výzvou pro webové archivy jsou pak sociální sítě, zejména např. Facebook nebo Twitter, ať už z výše jmenovaných hledisek nebo z pohledu vysoké poptávky, ze strany paměťových institucí po archivaci těchto zdrojů.

---

[85] AJAX je obecné označení pro technologie vývoje interaktivních webových aplikací.

[86] Interpret je speciální počítačový program, který umožňuje přímo vykonávat zápis jiného programu bez jeho převádění do strojového kódu.

[87] FARRELL, Susan a Kevin ASHLEY. *A guide to Web preservation: practical advice for web and records managers based on best practices from the JISC-funded PoWR project* [online]. S.l.: UKOLN / ULCC, 2010 [cit. 2015-03-03]. ISBN 09-516-8567-8. Dostupné z: http://jiscpowr.jiscinvolve.org/wp/files/2010/06/Guide-2010-final.pdf

Například sociální síť Twitter není tvořena jen příspěvky uživatelů, ale i jejich následnými konverzacemi. Je tedy nutné archivovat nejen příspěvky, ale i odpovědi na ně. Na druhou stranu odpovědi, které se archivují, nemusejí dávat smysl bez kontextu předchozích příspěvků.[88] Proto musí být zachovány při archivaci vazby mezi příspěvky. Veškerý obsah přibývá každou vteřinu a uživatelé mohou svoje příspěvky zpětně mazat. Součástí příspěvků uživatelů mohou být také odkazy mimo samotný Twitter, je pak otázkou, jestli by měl archiv mít obsah z těchto odkazů a jakým způsobem může zajistit časovou konzistenci mezi příspěvky a odkazy s ohledem na krátkou životnost odkazů sdílených na Twitteru.[89]

Na kompletním archivu sociální sítě Twitter již od roku 2010 pracuje Kongresová knihovna v USA (Library of Congress, Washington, D.C.). Kongresová knihovna uzavřela smlouvu přímo s Twitterem a získala od něj veškerá veřejná data od jeho založení až po rok 2010 a dohodu, že Twitter bude pravidelně dodávat nová data.[90] V polovině roku 2015 bylo v archivu Kongresové knihovny uloženo přes půl bilionu tweetů, ale knihovna zároveň začala mít s udržením archivu problémy.[91]

Kongresové knihovně se podařilo, díky dohodě s Twitterem, vyvarovat problémům, které akvizici webu 2.0 provází. Nicméně stále se musí potýkat se specifickými problémy dlouhodobé ochrany webu 2.0, a to s enormním objemem dat, jejich komplexitou a neustávajícím nárůstem. Pro dlouhodobou ochranu je zde největší výzvou právě zachování komplexity dat a zvolení vhodného formátu pro jejich uložení, který by nejen umožňoval jejich indexaci a vyhledávání, ale i splňoval kritéria pro souborový formát vhodný pro dlouhodobou ochranu. Přesně s těmito problémy se potýká i Kongresová knihovna a doposud (polovina roku 2016) svůj archiv nezpřístupnila veřejnosti.

---

[88] PENNOCK, Maureen. Web-Archiving. *DPC Technology Watch Series* [online]. 2013, roč. 13, č. 01 [cit. 2015-03-03]. DOI: http://dx.doi.org/10.7207/twr13-01. Dostupné z: http://dx.doi.org/10.7207/twr13-01

[89] LONG, Andrew Stawowczyk. *Long-term preservation of web archives: Experimenting with emulation and migration methodologies* [online]. Australia: National Library of Australia, 2009 [cit. 2015-03-03]. Dostupné z: http://www.netpreserve.org/sites/default/files/resources/Methodologies.pdf

[90] Update on the Twitter Archive At the Library of Congress. Library of Congress [online]. Washington, D.C.: Library of Congress, 2013 [cit. 2016-07-26]. Dostupné z: https://www.loc.gov/today/pr/2013/files/twitter_report_2013jan.pdf

[91] KLEIN, Lauren. An Archive of Tweets? In: Media, Materiality, and Archives [online]. Atlanta, Georgia, 2016 [cit. 2016-07-26]. Dostupné z: http://blogs.iac.gatech.edu/archives16/2016/01/26/an-archive-of-tweets/

Závěrem k této části věnované nástrojům je potřeba zmínit, že v současné době neexistuje univerzální nástroj, který by dokázal pokrýt celou agendu dlouhodobé ochrany webového obsahu. Většinou se jedná o aplikace, které byly vytvořeny na klíč specifickým potřebám jednotlivých webových archivů. Tyto aplikace pak bývají zapojovány do větších ekosystémů, které si instituce budují pro ochranu digitálního obsahu, ať už jen pro data z webových archivů nebo pro jejich celý digitální fond.

5. Příklad z praxe: Ukládání obsahu z Webarchivu do LTP úložiště NK ČR

Poslední část diplomové práce se zaměřuje na současnou praxi v České republice. Cílem je ukázat, jak dnes vypadá dlouhodobá ochrana českého webu. V této části bude velmi detailně popsán proces uložení dat z Webarchivu do LTP systému NK ČR. V předchozí části bylo ukázáno, jak vypadá tato praxe v zemích, kde má dlouhodobá archivace delší tradici a hlavně je podporována společností, která ji chápe jako nezbytnost pro zachování kulturního dědictví. S tím je samozřejmě spojeno i financování nejen samotných archivů, ale i výzkumu v této oblasti.

Ačkoliv v České republice nemá dlouhodobá archivace zdaleka takovou podporu, a to ani z řad některých členů knihovnické komunity, přesto se díky odborníkům působícím v tomto oboru podařilo, *„(…) vydobýt celosvětové uznání svými dlouhodobými aktivitami v oblasti digitalizace a digitální ochrany: v roce 2005 obdržela NK ČR cenu UNESCO/ JIKJI Memory of the World za svůj přínos v ochraně a zpřístupňování kulturního dědictví.“*[92] Velkým úspěchem bylo také, že se NK ČR, díky podpoře Ministerstva kultury, podařilo zajistit financování v rámci Integrovaného Operačního Programu (IOP) ze strukturálních fondů Evropské unie. A tak mohl vzniknout projekt nazvaný Vytvoření Národní digitální knihovny. *„V rámci tohoto dotačního projektu Národní knihovna České republiky a Moravská zemská knihovna v Brně zdigitalizují, dlouhodobě ochrání a zpřístupní významnou část svých fondů.“*[93] Národní digitální knihovna představuje velký milník pro českou dlouhodobou ochranu digitálních dokumentů, a to díky její velikosti a komplexnosti.

Pro účely této práce je důležité, že jedním z fondů, který má být dlouhodobě ochráněn a zpřístupněn, je fond webového archivu nazvaného Webarchiv (dříve také WebArchiv). Celá koncepce uložení dat z Webarchivu vznikala na půdě NK ČR a lze ji rozdělit na tři základní části:

1. Pre-ingest: příprava dat na straně oddělení webové archivace
2. Ingest (zpracování dat na straně LTP úložiště)
3. Správa dat v LTP systému

---

[92] HUTAŘ, Jan, Marek MELICHAR a Bohdana STOKLASOVÁ. Národní digitální knihovna. *Knihovna*. 2009, roč. 20, č. 1. Dostupné z: http://knihovna.nkp.cz/knihovna91/humesto.htm

[93] SVOBODA, Tomáš. Projekt Národní digitální knihovna: aktuální stav projektu. In: *INFORUM 2012: 18. konference o profesionálních informačních zdrojích*. Praha, 2012, s. 1-5. Dostupné z: http://www.inforum.cz/pdf/2012/svoboda-tomas.pdf

5.1. Východiska

Předtím, než se bude možné věnovat třem fázím koncepce uložení dat, je třeba se zaměřit na východiska, se kterými bylo nutné při její přípravě počítat. Základním východiskem bylo nastavení samotného Webarchivu.

Data z Webarchivu byla ukládána ve formátech ARC do poloviny roku 2013, každý o velikosti přibližně 100 MB[94]. Od půlky roku 2013 byl aktualizován sklízecí stroj a začalo se se sklízením do formátu WARC, velikost jednoho WARC souboru byla navýšena na 1 GB, protože předchozí limit s vývojem technologií již neměl opodstatnění. V současné době se počítá se dvěma sklizněmi celého českého webu ročně a s dalšími menšími výběrovými sklizněmi. Vždy po ukončení sklizně se budou data ukládat do LTP systému a do subsystému zpřístupnění. Subsystém zpřístupnění je modul, který má za úkol zpřístupňovat archivovaná data koncovým uživatelům. Tento modul není v rámci diplomové práce akcentován, neboť pro její účely není příliš podstatný.

Dalším zásadním faktorem při tvorbě koncepce je nastavení sklízecího stroje, neboť to určuje formu sklízení, která má dopad na výslednou podobu dat. Formou sklízení je myšleno, jakým způsobem jsou ukládány jednotlivé domény, resp. webové stránky do kontejnerových formátů. Webarchiv při sklízení nijak logicky nerozděluje sklizený webový obsah, protože sklízení probíhá paralelně, na rozdíl od sériového sklízení, které upřednostňuje Britská národní knihovna. U paralelního sklízení několik sklízecích strojů stahuje obsah z různých kanálů a ukládá ho do jednoho kontejneru, dokud není naplněna jeho kapacita. Po naplnění se kontejner uzavře a začne se plnit nový. Jeden kontejner tedy může obsahovat různé fragmenty z více domén, resp. webových stránek. Tento přístup přináší větší efektivitu při jejich sklízení.

[94] Limit velikosti balíčku nastavený pro sklízecí stroj.

5.2. Intelektuální entita

Intelektuální entita reprezentuje míru nastavení granularity. „*Granularita určuje, co je základní entitou informačních toků a procesů v daném kontextu, z jakého bodu lze dále seskupovat a strukturovat vyšší celky a jaké jsou možnosti pro kombinace a konfigurace základních entit.*“[95] V počátku tvorby koncepce bylo nutné vyřešit otázku, do jaké hloubky se bude s webovým obsahem pracovat. Analogicky mají například knihovny zvolenou jako svoji entitu typicky jeden svazek knihy u monografií nebo jedno číslo u periodik. Se zvolenou entitou pak nadále pracují: identifikují ji, zkatalogizují a pak nad těmito entitami vyhledávají. Tento příklad je z analogového světa, v případě digitálního prostředí je situace daleko problematičtější, přesto je typické, že „*v digitálním světě je základní jednotkou pro identifikaci a správu v rámci operačních systémů zpravidla jeden počítačový soubor*“[96].

V případě tradičních zdigitalizovaných dokumentů sahají digitální knihovny ke stejnému principu jako u analogových dokumentů, tedy u monografií bude intelektuální entitou jeden svazek předlohy, přestože zdigitalizovaná monografie může být tvořena více počítačovými soubory.

U webového obsahu je zvolení intelektuální entity zásadním rozhodnutím, neboť to ovlivní nejen přístup a správu samotných dat, ale navíc není možné hledat inspiraci v tradičním knihovnictví, tak jako u digitalizovaných dokumentů. Definování entity také přímo ovlivňuje možnou míru digitální ochrany. Tím, do jaké hloubky bude definována granularita, se poté určuje míru kontroly nad daty, možnosti plánování ochrany a s tím spojenou náročnost na výpočetní výkon nebo diskové kapacity.

Jako první se nabízelo řešení použít jako intelektuální entitu jednu doménu nebo webový server analogicky k jednomu svazku knihy, neboť je člověk (čtenář) apriori vnímá jako celistvý objekt. Internetovou stránku uživatel bude hledat a právě s tímto objektem chce pracovat. Vzhledem ke zvolené formě sklízení (kap. 5.1.) by bylo takové řešení přespříliš náročné na výpočetní výkon, neboť by se obsah kontejnerů musel přeorganizovat. Navíc kvůli rizikům spojeným s přesunem počítačových souborů a s tím spojenou transformací metadat bylo toto řešení nutné vyloučit.

---

[95] CUBR, Ladislav. *Dlouhodobá ochrana digitálních dokumentů*. 1. vyd. Praha: Národní knihovna České republiky, 2010, 154 s. ISBN 978-80-7050-588-5.

[96] Tamtéž jako 94.

Stejně tak bylo nutné vyloučit jednotlivé počítačové soubory jako intelektuální entity, zejména kvůli náročnosti na výpočetní výkon, neboť se při takovém řešení zvyšuje řádově počet entit (webové stránky se mohou skládat ze stovek či tisíců jednotlivých počítačových souborů). Metadatový popis takových entit by příliš nepřesahoval metadata uložená v hlavičce WARCu a problémem je i jednotlivá provázanost souborů (jedna HTML stránka může být tvořena z několika souborů se skripty).

Z toho důvodu byl základní intelektuální entitou zvolen jeden archivní kontejner (soubory sklizené z webové stránky zabalené do kontejnerového formátu), se kterým se bude pracovat v rámci ochrany jen na jeho úroveň. V tuto chvíli bude samotný obsah uvnitř kontejneru ignorován, tzn. v metadatech nebudou nijak popisovány soubory uvnitř kontejneru.

Důvody, které vedly k tomu rozhodnutí:

- Velká obsáhlost kontejnerů. Jeden kontejner může obsahovat řádově tisíce souborů v nejrůznějších formátech.
- Metadata k souborům jsou obsažena přímo v samotném kontejneru (crawler Heritrix vkládá ke sklizeným souborům technická a administrativní metadata viz kap. 5.5.3.).
- Jak již bylo zmíněno výše. Obsah jednoho www serveru/domény je roztroušen do více kontejnerů, tzn. jen kontejner nerovná se jeden www server/doména. Jeden kontejner obsahuje různé fragmenty z různých www serverů/domén.

Nicméně pouze kontejner jako intelektuální entita není dostačující. Vzhledem k tomu, že nelze jednoduše určit, kde jsou jednotlivé komponenty serverů/domén uloženy a ani jaké kontejnery jsou navzájem propojeny, je nutné vytvořit ještě sekundární intelektuální entitu, tzv. sklizeň. Pomocí metadat sklizně lze určit, které kontejnery spolu souvisejí, a mít tak pohromadě veškerá potřebná data k zobrazení sklízených serverů. Bylo potřeba zohlednit sklizeň jako jistou formu logické jednotky, která bude „svoje“ kontejnery zastřešovat. Proto vznikly dvě různé intelektuální entity, které jsou vzájemně propojené. Primární entitou je jeden kontejner, který obsahuje vlastní data. A pak je tu zastřešující sekundární entita sklizně, která se skládá ze souborů s nastavením sklízecího stroje, logy a reporty.

5.2.1. Soubory s nastavením, logy a reporty

Soubory s nastavením, logy a reporty jsou textové soubory, které generuje sklízecí stroj při procesu sklízení. Informace obsažené v těchto souborech se vztahují k celému procesu sklízení dat. Díky nim je možné zpětně rekonstruovat události při průběhu sklizně. Konkrétně pak obsahují informace k událostem během sklizně, jako je seznam sklizených souborů, jejich velikost, nebo třeba záznamy o souborech, které se nepodařilo sklidit. Dále obsahují kompletní nastavení sklízecího stroje, dobu trvání a velikost sklizně. Některé informace jsou duplikovány a lze je zpětně získat ze samotných dat, ale zejména informace o chybách, stavy stahování jednotlivých souborů jsou zapsány jen v těch souborech a již je nelze zpětně vygenerovat.

Velikost souborů s nastavením, logy a reporty je odvozena od velikosti sklizně, u standardních sklizní se jejich objem pohybuje v řádech stovek MB až jednotek GB. Celkově jsou soubory natolik obsáhlé, že s nimi nelze jednoduše pracovat. Proto není možné je přidat k metadatovým souborům, které budou vytvářeny k jednotlivým kontejnerům. Z tohoto důvodu je nelze mít uložené přímo v balíčku s daty. Soubory s nastavením, logy a reporty jsou jedním z důvodů, proč musela vzniknout sekundární entita sklizně.

5.2.2. Struktury uložení – archivní balíček NDK

Veškerý obsah, který je ukládán v LTP úložišti, je distribuován v tzv. balíčcích (jak to ukládá norma OAIS). V LTP úložišti NK ČR jeden balíček obsahuje jednu intelektuální entitu, v případě webového obsahu pak jeden kontejner s daty. Balíčky mění svůj typ v závislosti na tom, na jakém místě systému se právě nachází.

Celé LTP úložiště NK ČR vychází z požadavků na důvěryhodný digitální depozitář daný konceptuálním modelem normy OAIS, z této normy také vychází typy a názvosloví balíčků. Jednotlivé typy balíčků, které se v normě objevují, jsou:[97]

- SIP = Submission information package – balíček dat a metadat vstupující do LTP úložiště
- DIP = Dissemination information package – balíček dat a metadat vystupující z LTP úložiště
- AIP = Archival information package – balíček dat a metadat na LTP úložišti

[97] HUTAŘ, Jan. Podrobnější popis projektu NDK a jeho kontext. NÁRODNÍ KNIHOVNA ČESKÉ REPUBLIKY. *Národní digitální knihovna* [online]. 2010 [cit. 2015-09-19]. Dostupné z: https://web.archive.org/web/20110106000746/http://ndk.cz/narodni-dk/podrobnejsi-popis-projektu/podrobnejsi-popis-projektu-ndk

Výše uvedené balíčky se mohou lišit svým obsahem, ale nemusí, např. u LTP úložiště NK ČR se archivní balíček rovná balíčku určenému pro export, ale na druhou stranu SIP balíček je transformován a dojde k jeho obohacení o nové metadatové soubory dříve, než se stane archivním balíčkem. V diplomové práci se pracuje pouze s balíčky typu SIP a AIP, neboť DIP balíček má stejnou podobu jako AIP.

Vzhledem k tomu, že existují dvě různé intelektuální entity pro webový obsah, musí být definované dva různé AIP balíčky, jeden pro kontejner a jeden pro sklizeň. Oba balíčky vychází z obecné definice balíčků, kterou používá LTP úložiště NDK. Pro lepší ilustraci je níže uvedena struktura balení dat a metadat v jednom AIP balíčku monografického dokumentu (tabulka č. 1: Struktura balíčku s monografií[98]).

Tabulka 1: Struktura balíčku s monografií

ADRESÁŘ>	OBSAHUJE>>	OBSAHUJE>>>
Monografie	info.xml	
	masterCopy(adresář)	obrazy JPEG 2000 lossless
	ALTO (adresář)	soubory ALTO.xml pro každou stranu
	amdSec (adresář)	AMD_METS.xml pro každou stranu
	hlavní_METS.xml	
	soubor.md5	

V tabulce je naznačena adresářová struktura balíčku jednoho svazku digitalizovaného monografického dokumentu. Celý balíček je tvořen hierarchickou strukturou adresářů, ve které jsou uloženy metadatové soubory a obrazová data.

V kořenovém adresáři jsou umístěny hlavní metadatové informace vztahující se k celému balíčku, resp. celému dokumentu. Soubor *hlavní_METS.xml* obsahuje popisná metadata dokumentu, v tomto případě se jedná o bibliografické informace o publikaci. Dále pak obsahuje strukturální metadata, která obsahují mapu souborů, a spojují tak jednotlivé soubory v balíčku do kompletního celku, tzn. pomocí těchto metadat lze zpětně zrekonstruovat digitalizovanou publikaci. Soubor *.md5* obsahuje kontrolní součet pro každý soubor v balíčku, tím je zajištěna bitstreamová ochrana souborů. A jako

[98] ŠVÁSTOVÁ, Pavla a Jaroslav KVASNICA. Definice metadatových formátů pro digitalizaci monografických dokumentů (monografií, kartografických dokumentů, hudebnin). NÁRODNÍ KNIHOVNA ČESKÉ REPUBLIKY. Národní digitální knihovna [online]. 2013 [cit. 2015-09-15]. Dostupné z: http://www.ndk.cz/archivace/DMF-monografie-1-1.pdf

poslední je v kořenovém adresáři soubor *info.xml*, který má v sobě obsažené informace o vzniku balíčku.

V jednotlivých adresářích jsou rozmístěna data, textové soubory a metadata – v adresáři *masterCopy* digitalizované stránky, v *ALTO* výstupy z jejich OCR[99] a v *amdSec* technická a administrativní metadata pro jednotlivé stránky.

5.2.3. Struktury uložení – archivní balíček pro kontejnerový formát

Obdobně jako u balíčku s monografií vypadá i struktura balíčku pro kontejnerový formát (tabulka č. 2: struktura balíčku s kontejnerem[100]). Kořenový adresář obsahuje: soubor se strukturálními a popisnými metadaty, soubor se základními informacemi o vzniku balíčku a také soubor s MD5 kontrolními součty.

Tabulka 2: struktura balíčku s kontejnerem

ADRESÁŘ>	OBSAHUJE>>	OBSAHUJE>>>
ARC/WARC	info.xml	
	data(adresář)	data v kontejnerovém formátu
	TXT (adresář)	extrahované textové soubory
	amdSec (adresář)	metadatové soubory
	hlavní_METS.xml	
	soubor.md5	

V adresáři *data* je umístěn kontejner se samotnými daty. Adresář *TXT* obsahuje extrahované textové soubory, pokud existují, je to tedy nepovinná součást balíčku. Adresář *amdSec* má v sobě soubor s technickými a administrativními metadaty. Přestože adresáře *data* a *amdSec* obsahují pouze po jednom souboru, jsou rozmístěny v adresářích kvůli předpokladu, že postupem času budou vznikat jejich další kopie, např. při formátové migraci.

5.2.4. Struktury uložení – archivní balíček pro sklizeň

Struktura balíčku pro sklizeň je jednodušší než ta pro data. Obsahuje pouze jeden adresář, ve kterém jsou uloženy logy, reporty a soubory s nastavením. V kořenovém adresáři je opět trojice souborů s metadatata a MD5 kontrolními součty. Jak je patrné ze

[99] OCR je optické rozpoznávání znaků je metoda pro digitalizaci tištěných textů

[100] KVASNICA, Jaroslav a Rudolf KREIBICH. Specifikace pro data z WA: pouze pro formát WARC. Praha, 2014.

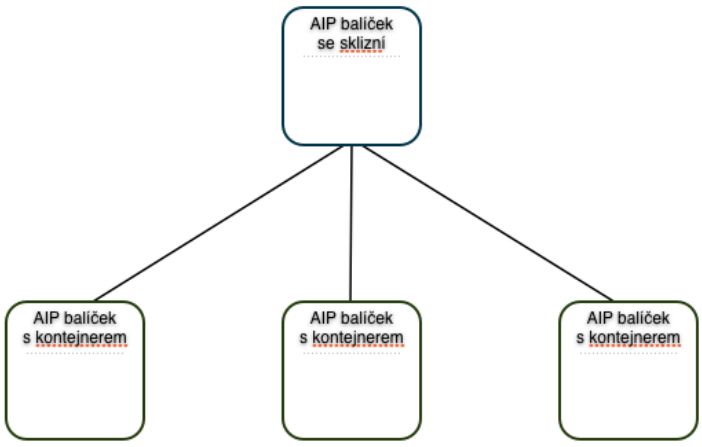
struktury (tabulka č.3: struktura balíčku pro sklizeň[101]), tak balíček sklizně neobsahuje žádná vlastní data z webu, slouží jen jako zastřešovací element pro data uložená v balíčcích pro kontejnery.

Propojení mezi balíčky s daty a balíčkem pro sklizeň je realizováno pomocí identifikátorů. V metadatech v archivním balíčku pro sklizeň je uložený kompletní seznam identifikátorů balíčků s kontejnery. A každý kontejnerový balíček má v metadatech zapsaný jedinečný identifikátor sklizně, ke které patří. Tím je zajištěno oboustranné propojení dat se sklizní.

Tabulka 3: struktura balíčku pro sklizeň

ADRESÁŘ>	OBSAHUJE>>	OBSAHUJE>>>
Sklizeň	info.xml	
	data(adresář)	soubory s logy, reporty a nastavením
	hlavní_METS.xml	
	soubor.md5	

V praxi pak vzniká jeden balíček sklizně a několik stovek až tisíců balíčků s kontejnery, dle velikosti sklizně. Například pro sklizeň velkou 1TB bude jeden balíček sklizně zastřešovat cca deset tisíc[102] balíčků s kontejnery. Celý seznam všech balíčků je uložen v souboru hlavní_METS.xml u balíčku sklizně. Na následujícím obrázku (obrázek č. 2 hierarchická struktura balíčků) je graficky znázorněno toto propojení, které připomíná hierarchický strom pouze s jednou úrovní.



Obrázek č. 2 hierarchická struktura balíčků

[101] KVASNICA, Jaroslav a Rudolf KREIBICH. Specifikace pro data z WA: pouze pro formát WARC. Praha, 2014.

[102] V případě starých typů sklizní s velikostí balíčku 100MB

Jak již bylo zmíněno v úvodu této části, než se data dostanou do úložiště, musejí projít zpracováním. První fází zpracování je tzv. pre-ingest, druhou fází je ingest a třetí představuje samotné uložení v LTP úložišti.

5.3. Pre-ingest

Fáze pre-ingest představuje základní přípravu a formalizaci dat tak, aby mohla být přesunuta k další fázi ingestu. Ingestem je myšlena transformace a validace dat do formátu, se kterým pracuje LTP úložiště. Pre-ingest probíhá ještě na straně Webarchivu, data jsou stále uložena na jeho úložišti a využívá se jeho výpočetní výkon. Celý pre-ingest spočívá v doplnění samotných dat o informace, které obsahují další systémy, s nimiž pracují kurátoři českého webového archivu.

Kurátoři Webarchivu využívají ke správě dat aplikaci WA Admin, která byla vyvinutá přímo na míru tomuto účelu. Obsahuje evidenci zdrojů, které jsou součástí tzv. výběrových sklizní, a webový archiv je v kontaktu s jejich vydavateli. Ve WA Adminu jsou obsažena bibliografická metadata k jednotlivým zdrojům, jejich kurátorská hodnocení a evidence smlouvy o přidělení výhradní licence pro veřejné zpřístupnění zdroje. Jsou zde uchovávány informace k poměrně malé části všech sklizených webových stránek, protože není v lidských silách zkatalogizovat kompletně český web. Přesto jsou tyto informace velmi cenné, a proto jsou přidávány ke sklizeným datům před uložením do archivu.

Základní nástroj pro pre-ingest je Software pro automatizovaný metadatový popis sklizní Webarchivu (dále jen SAMPSA), který extrahuje informace ze tří různých zdrojů: z dat v kontejnerových formátech, z kurátorské aplikace WA Admin a z logů, reportů a souborů s nastavením. Tato data transformuje a ukládá je jako standardizovaný XML dokument, který pak slouží pro obohacení metadatového záznamu v LTP úložišti NK ČR.

Základní funkcí SAMPSA je schopnost automatizovaně pracovat s uvedenými zdroji. Umí si otevřít zdrojové soubory v kontejnerových formátech a vyextrahovat z nich potřebná metadata. A to jak pro formát WARC, tak i pro starší formát ARC. SAMPSA prochází adresářovou strukturu v úložišti archivu, prochází jednotlivé kontejnery ve sklizni a z nich extrahuje metadata. Zároveň zapisuje jednotlivé kontejnery do logické strukturální mapy, která je součástí metadatového popisu, a tím vzniká kompletní seznam kontejnerů patřících ke sklizni.

Z logů, reportů a souborů s nastavením a propojením s metadaty ve WA Adminu vytvoří aplikace SAMPSA bibliografický popis sklizně, podrobněji o metadatech v kap. 5.5.

Další funkcí, kterou SAMPSA má na starosti, je přidělování identifikátoru, který je jedinečný v rámci LTP úložiště. Jako identifikátor je používán UUID, který je generován náhodně, ale zároveň je ukládán do databáze, aby byla zajištěna jeho unikátnost. Po provedení všech výše zmíněných kroků pak aplikace uloží veškerá data jako strukturovaný standardizovaný XML soubor, který je přiložen ke sklizni u logů, reportů a souborů s nastavením.

Jedním z dalších procesů pre-ingestu je vytváření indexu webových stránek. K vytváření indexu je využívána softwarová knihovna, která je součástí aplikace Wayback Machine. Index, který je vytvářen v části pre-ingestu, obsahuje kompletní výpis URL adres dané sklizně, ze které pocházejí stažené soubory.

S vytvořeným indexem pak pracuje další aplikace, která obohacuje metadata v balíčku, o metadata z katalogizačních záznamů v centrálním katalogu NK ČR. Část webových zdrojů je katalogizována a má vytvořený kompletní bibliografický popis, jako ostatní dokumenty z fondu NK ČR. Aplikace vyhledává zkatalogizované webové zdroje, které se nacházejí v datech, která byla sklizena při archivování českého webu.

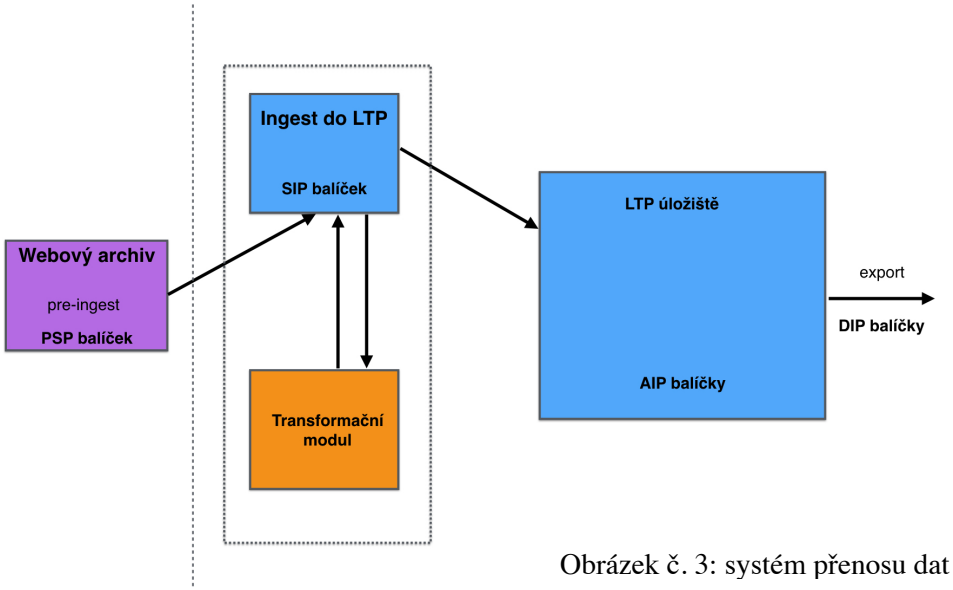
V centrálním katalogu je využíván pro katalogizaci formát MARC, který ale není vhodným formátem pro dlouhodobou ochranu, proto jsou nalezené záznamy transformovány do metadatového standardu MODS, který je postaven na značkovacím jazyku XML. V případě, že je nalezena shoda, je pro stažení metadatového záznamu z centrálního katalogu NK ČR využíván protokol OAI-PMH[103] a následná transformace je realizována pomocí, pro české prostředí upravené šablony, MARCtoMODS[104] od Kongresové knihovny. Po transformaci je záznam uložen jako XML dokument do adresáře *dmddata*, který je vytvořen na stejném místě, kde jsou uloženy reporty, logy a soubory s nastavením.

[103] OAI-PMH je protokol pro sběr metadatových záznamů

[104] MARCtoMODS je šablona pro převod metadatových formátů MARC do MODS

5.4. Ingest

Po dokončení posledního kroku pre-ingestu se celá sklizeň, včetně metadatových souborů, přesouvá na dočasné úložiště LTP systému NK ČR, kde k ní má přístup transformační modul. Transformační modul je rozšiřitelná funkční komponenta, která má v LTP systému NK ČR na starost veškeré procesy spouštěné nad daty. V momentě, kdy webový archiv předává SIP balíček, se data stávají součástí LTP systému NK ČR. Průchod balíčku je znázorněn na obrázku níže (obrázek č. 3: systém přenosu dat), kdy černé šipky naznačují tok dat, modré čtverce datové úložiště, oranžový čtverec transformační modul a fialový čtverec pak interní úložiště webového archivu.

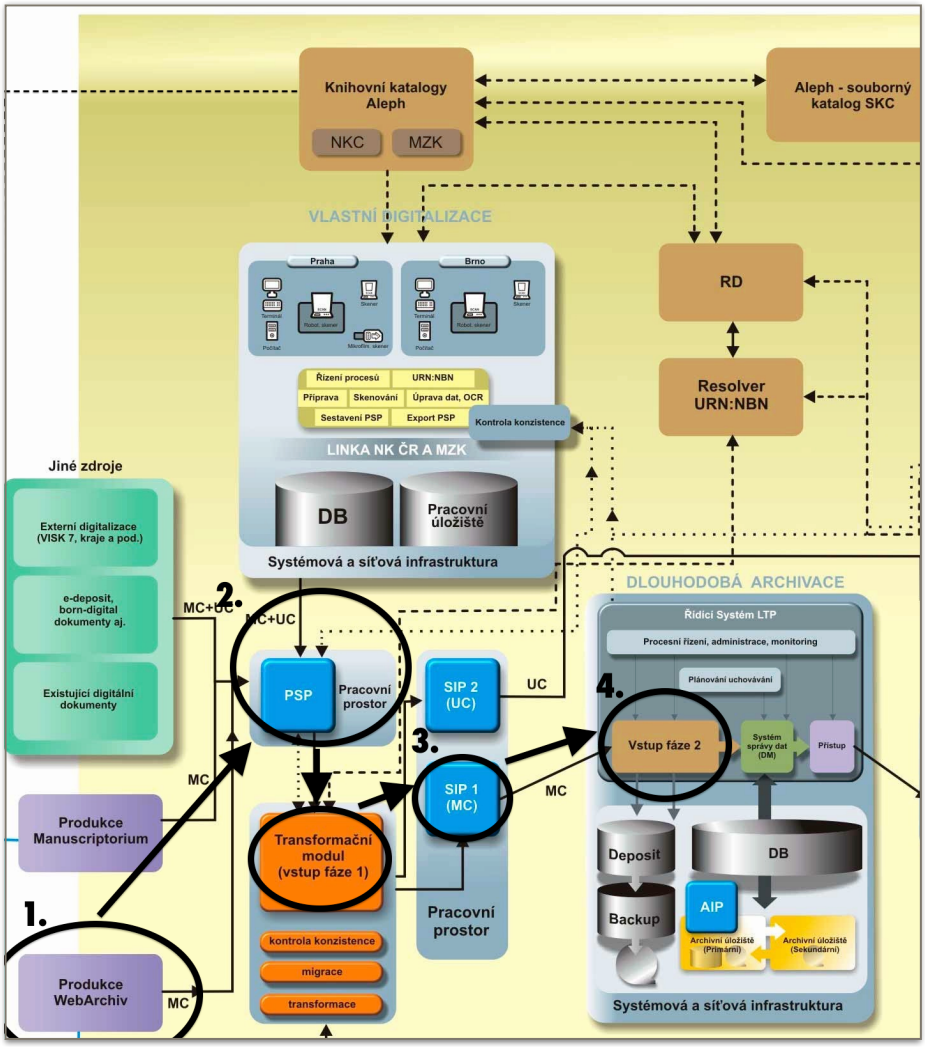


Obrázek č. 3: systém přenosu dat

Z výše uvedeného vyplývá, že v pre-ingestu se vytváří SIP balíček, který si po přesunu na dočasné úložiště převezme transformační modul a vytvoří z něj AIP balíček, který je uložen v LTP úložišti. Pro úplnost je ve schématu naznačen i export dat, ten probíhá ve formě DIP balíčků, které jsou vytvořeny z AIP balíčků a odeslány mimo LTP systém.

Na dalším obrázku (obrázek č. 4: schéma Národní digitální knihovny[105]) je naznačen průchod dat prostřednictvím komplexního schématu celého systému Národní digitální knihovny, který se stará o digitální dokumenty NK ČR a jehož součástí je i webový archiv a LTP úložiště, dalšími součástmi jsou pak např. digitalizační linka nebo resolver identifikátoru URN:NBN.

[105] Podrobnější popis projektu NDK a jeho kontext. Portál Národní digitální knihovny [online]. Praha, 2011 [cit. 2016-07-26]. Dostupné z: http://wayback.webarchiv.cz/wayback/20140320103556/http://ndk.cz/narodni-dk/podrobnejsi-popis-projektu



Obrázek č. 4: schéma Národní digitální knihovny

LTP úložiště může přijímat data z různých zdrojů v různých formátech. Jedním ze zdrojů připojených k systému jsou data z webového archivu (na schématu: Produkce WebArchiv, 1.), která jsou z něj odesílána do pracovního prostoru (na schématu: PSP, pracovní prostor, 2.). V pracovním prostoru s nimi pracuje transformační modul, který provádí transformace dat, tvoří metadata a kontroluje jejich validitu a konzistenci.

Ve schématu je naznačen nový typ balíčku PSP (Producer Submission Package), ten byl definován v začátku projektu NDK skupinou odborníků, kteří chtěli odlišit balíček z externí digitalizace a balíček připravený přímo pro import do LTP systému. PSP balíček není součástí referenčního modelu OAIS a v dnešní době se ukazuje, že existence dalšího balíčku pro producenty dat nebylo opodstatněné. V praxi se prokázalo, že definování dvou různých balíčků pro vstup je zbytečným mezikrokem, a v současné době je PSP balíček stejný jako SIP balíček (na schématu: Pracovní prostor 2. a 3. jsou totéž).

Při ukládání dat na úložiště je provedena jejich finální kontrolní validace (na schématu: Vstup fáze 2, 4.) a následně jsou data v AIP balíčku uložena na fyzické nosiče.

To, jak vypadá SIP balíček (a zároveň PSP balíček) pro webový obsah, tedy jakou mají mít data a metadata strukturu, aby mohla být přijata do LTP systému, je definováno ve Specifikaci pro data z webového archivu[106]. Specifikace předepisuje podobu SIP balíčku a zároveň AIP balíčku a DIP balíčku. Specifikace jsou celkem dvě, jedna pro balíčky s ARC kontejnery (Příloha 1) a jedna pro balíčky s WARC kontejnery (Příloha 2). Příklady definice struktury balíčků v textu pochází pouze ze specifikace pro WARC formát, protože se obě specifikace od sebe liší jen v detailech. Kompletní specifikace jsou připojeny k diplomové práci jako přílohy.

SIP balíček je definován adresářovou strukturou, označením mandatorních adresářů a očekávanými souborovými formáty v nich:

```
/
*.warc.gz
*.warc.gz.open -- bude nahrán stejně jako formát warc.gz
./logs
        *harvest.xml [mandatory]
        ./dmdsec
             *.xml
        ./index [mandatory]
             *.cdx
       ./crawl [mandatory]
             *.tar.gz
```

Při ingestu do LTP systému je posílán jeden SIP balíček, který je reprezentován jedním adresářem, ve kterém jsou nahrány všechny kontejnery patřící ke vkládané sklizni. Dále tento adresář musí obsahovat metadatový soubor harvest.xml a podadresáře index a crawl. Obsah souboru harvest.xml bude podrobně rozebrán v následujících kapitolách. V podadresářích jsou uloženy soubory s indexem a logy, reporty a soubory s nastavením.

---

[106] KVASNICA, Jaroslav a Rudolf KREIBICH. Specifikace pro data z WA: pouze pro formát WARC. Praha, 2014.

Další adresář, který může a nemusí být v SIP balíčku přítomen, je *dmdsec*. Obsahuje bibliografická metadata stažená z katalogu Aleph.

5.4.1. Transformační modul

Iniciace importu dat do LTP úložiště probíhá pomocí grafického rozhraní aplikace zvané Workflow, která má na starosti správu příjmu dat do LTP úložiště. V tomto rozhraní technický správce vidí i stav importu a případná chybová hlášení.

Po iniciaci importu transformační modul nejprve vytvoří balíček se sekundární intelektuální entitou sklizně a ke sklizni přidělí unikátní identifikátor. Jako identifikátor je využíván UUID, který byl vygenerován již v pre-ingestu, a je uložen v souborech s nastavením, logy a reporty (konkrétně v souboru harvest.xml). Tento identifikátor je pak vkládán k balíčkům se samotnými daty, tím je zajištěna jejich příslušnost ke své sklizni. Při tvorbě balíčku vytvoří transformační modul i adresářovou strukturu, která je popsána v předchozí kapitole, a uloží soubory s nastavením, logy a reporty.

Po vygenerování XML souborů a vytvoření MD5 souboru transformační modul provede validaci všech vygenerovaných dat. Validace kontroluje, jestli obsahuje výsledný balíček všechny povinné soubory a jestli metadata obsahují všechny povinné položky. Takto zkompletovaný balíček pak transformační modul posílá do repozitáře. Před přijmutím repozitáře probíhá ještě jedna kontrolní validace ze strany repozitáře.

Jakmile v pořádku doputuje do LTP systému balíček se sklizní, pak začíná postupný import všech archivních kontejnerů. Jelikož jsou takových balíčků řádově stovky až tisíce, tak import souborů z ostatních zdrojů má přednost a import dat z webového archivu probíhá téměř výhradně v nočních hodinách, kdy je LTP systém nejméně vytížen. Stejně jako u předchozího importu transformační modul generuje metadatové soubory, zároveň ale probíhají i procesy nad samotným archivním kontejnerem, ze kterého jsou extrahovány některé informace, které jsou pak uloženy do metadat.

5.4.2. Migrace kontejnerových formátů

Český webový archiv má v současné době data uložená v obou kontejnerových formátech (ARC i WARC). Důvodem je historický vývoj a stáří archivu, který vznikal již v roce 2001. V počátcích se začalo sklízet do staršího formátu ARC, neboť WARC ještě nebyl k dispozici. V současné době je větší část archivu ve formátu WARC, na který se přešlo v roce 2013. Nicméně ARC je zastaralý a nedokonalý formát a jeho využívanost u webových archivů klesá. Proto byl vyhodnocen jako rizikový pro dlouhodobou ochranu

dat. Na základě těchto faktů se dospělo k rozhodnutí, že před vstupem do LTP systému bude starší formát ARC migrovat do pokročilejšího formátu WARC.

Samotná migrace je také prováděna transformačním modulem, který využívá open-source nástroj warc-tools vyvíjený firmou Hanzo, která se zabývá archivováním webů pro komerční subjekty. Principiálně je migrace mezi dvěma kontejnery pouze doplněním metadatových hlaviček ke všem souborům v kontejneru. Bohužel, ne všechna metadata lze zpětně rekonstruovat, protože vznikají již ve chvíli, kdy jsou data sklízena. Jako u každé migrace mezi souborovými formáty je zde riziko chybovosti, proto je nutné provádět zpětně kontroly, jestli nebyly při migraci poškozeny soubory uvnitř kontejneru a zda je možné z migrovaných kontejnerů zpětně rekonstruovat webový obsah. Z toho důvodu zůstávají ARC kontejnery zachovány a jsou součástí AIP balíčku s nižším statusem digitální ochrany.

Dalším procesem, který může být prováděn nad samotnými daty, je extrahování textu. V současné době je tento proces, protože neúměrně zatěžuje výpočetní výkon a zpracování velkých sklizní zabere příliš mnoho času. Pro tento proces je využívána aplikace Apache Tikka. Transformační modul otevře všechny kontejnerové soubory a z objektů, které jsou v nich uloženy, vyextrahuje čistý text bez formátování. Text pak uloží do textových souborů s příponou .txt, které následně slouží pro fulltextové vyhledávání v obsahu, který je uložený v archivech. Fulltextové vyhledávání nad webovým archivem zvyšuje uživatelský komfort pro používání webového archivu, ale není nezbytné pro procesy dlouhodobé ochrany.

5.4.3. Řízení importu technickým správcem

Při vytváření balíčků s daty může dojít k situaci, kdy se některé balíčky nepodaří korektně vytvořit. Z tohoto důvodu musí dojít k částečnému řízení importu technickým správcem systému, který průběžně kontroluje, zda nedošlo k chybě. Celý proces je zpracováván po dávkách, které se postupně přijímají, a až poté dojde k rozdělení na jednotlivé entity, ze kterých jsou individuálně vytvářeny balíčky. Celé řízení spočívá v tom, že import je založen na principu tzv. pseudo-transakčnosti, kdy nejprve systém rozdělí data na balíčky a v případě chyb čeká na rozhodnutí správce, jak pokračovat, a teprve poté všechna data zpracuje.

Pokud transformační modul ohlásí, že se některý z balíčků nepodařilo vytvořit nebo z nějakého důvodu neprošel validacemi do dlouhodobého úložiště, pak musí dojít

k aktualizaci sklizně. To znamená, že transformační modul podle identifikátoru sklizně zjistí, zda je již založená, a pokud ano, tak provede kontrolu, jestli daný balíček již existuje a případně jej do sklizně doplní. Tento princip je důležitý, neboť ve chvíli, kdy dojde k uložení dat do repozitáře, není možné je již mazat nebo zaměňovat.

5.5. Metadatový popis

Při tvorbě metadatové struktury bylo vycházeno z Definice metadatových formátů pro digitalizaci monografických dokumentů (monografií, kartografických dokumentů, hudebnin)[107]. Za prvé definice metadatových formátů pro data z webového archivu zachovává logickou strukturu metadatových souborů v balíčku, a to jak pro primární, tak i sekundární entitu. Za druhé zachovává i metadatové standardy, které jsou používány při popisu monografických dokumentů.

Konkrétně se jedná o strukturální metadatový standard METS, který slouží jako formát pro vkládání ostatních metadatových standardů a definování struktury dokumentu. Dalším standardem je MODS pro popisná metadata a PREMIS pro technická a administrativní metadata. Stejně jako u monografických dokumentů jsou technická a administrativní metadata oddělená od popisných, tzn. jsou uložena v různých souborech.

Stejně jako se balíček vytváří postupně průchodem všemi fázemi jeho přípravy, tak i metadata vznikají na různých místech. V první fázi vznikají metadata již při sklízení webového obsahu a generuje je sklízecí stroj. Tato metadata jsou nazývána reporty, logy a soubory s nastavením a byla jim věnována kap. 5.2.1. Nicméně kvůli jejich robustnosti se tato metadata nepoužívají k samotné aktivní ochraně, tedy nejsou využívána jako zdroj informací pro ochranné aktivity. Jejich hlavní účel je uchovávat kompletní informace o původu a průběhu sklizně a zároveň jsou z nich čerpány informace do metadatových souborů, se kterými pracují kurátoři. Další nezpochybnitelnou hodnotu mají tato metadata pro badatele, kteří budou chtít nahlédnout do vývoje českého internetu.

5.5.1. Metadata ve fázi pre-ingestu

Ve fázi pre-ingestu vznikají metadata na straně Webarchivu. Ke generování metadat je používán software, který byl vyvinut speciálně pro tento účel. Jedná se zejména o uložení informací, které vznikají činností kurátorů Webarchivu. Kurátoři vybírají významné zdroje, které katalogizují a vedou si o nich záznamy. Tyto zdroje jsou pak sklízeny častěji a většinou probíhá komunikace s autorem stránek, který povolí veřejné zpřístupnění jeho archivovaných webových stránek uživatelům Webarchivu. Další možností jsou zdroje zpřístupněné pod licencí Creative Commons. Veškeré tyto

[107] ŠVÁSTOVÁ, Pavla a Jaroslav KVASNICA. Definice metadatových formátů pro digitalizaci monografických dokumentů (monografií, kartografických dokumentů, hudebnin). NÁRODNÍ KNIHOVNA ČESKÉ REPUBLIKY. Národní digitální knihovna [online]. 2013 [cit. 2015-09-15]. Dostupné z: http://www.ndk.cz/archivace/DMF-monografie-1-1.pdf

informace o zdrojích jsou staženy z kurátorských systémů a strukturovány do standardních metadatových formátů. V případě bibliografických metadat je používán standard MODS a pro strukturální metadata standard METS.

Mimo kurátorských dat je pro každý balíček sklizně generován soubor, který je nazýván *harvest.xml*. Tento soubor obsahuje základní informace o sklizni a hlavně její unikátní identifikátor, který je sklizni přidělen a slouží ke spárování dat se samotnou sklizní. Pro generování unikátního identifikátoru je využíván systém UUID[108] a zároveň na straně Webarchivu běží databáze s těmito identifikátory a je hlídána jejich unikátnost. V souboru je také uložena strukturální mapa ve formátu METS, která obsahuje kompletní seznam všech kontejnerových souborů, které patří pod sklizeň. Tento seznam slouží k jedné z kontrol, že importovaná sklizeň byla uložena kompletní.

```
<oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/">
 <dc:title>WARC-TEST-R18</dc:title>
 <dc:type>Crawl</dc:type>
 <dc:audience>ABA 001</dc:audience>
 <dc:identifier>a92r1g8x-84fp-45qt-la-9my53b5kfcdd</dc:identifier>
</oai_dc:dc>
```

Příklad výše slouží jako ukázka základních popisných metadat pro sklizeň, ve kterých je uložený název sklizně, typ sklizně, pro koho byla sklizeň vytvořena a unikátní identifikátor sklizně. Metadata jsou uložena ve standardu Dublin Core. S těmito metadaty pak dále pracuje transformační modul, který je převádí do formátů, které využívá LTP systém.

5.5.2. Metadata v LTP systému

Z definované struktury archivních balíčků plyne, že pro balíček se sklizní je generovaný jeden metadatový soubor a pro balíček se samotnými daty dva. Jak bylo zmíněno výše, metadata kopírují model, který je využíván u monografií nebo periodik. To znamená, že balíček s daty má jeden hlavní soubor, který obsahuje popisná metadata a strukturální mapy, a druhý soubor s technickými a administrativními metadaty, který obsahuje podrobné informace o datech.

---

[108] UUID (Universally unique identifier) je standard pro tvorbu identifikátorů. Je generován pomocí algoritmu.

Balíček sklizně, ve kterém nejsou uložena žádná data, obsahuje pouze jeden hlavní metadatový soubor se základním popisem a strukturální mapou. Na dalším příkladu je vidět zápis sekce *metsHdr*, která obsahuje informace o samotném metadatovém dokumentu. Dále tu je vytvoření dokumentu a jeho poslední modifikace, které jsou zde stejné, protože dokument nebyl zatím modifikován. Dále jsou v příkladu dva tzv. METS agenti. Jeden je pro tvůrce metadatového dokumentu a jeden pro roli archiváře, který má na starosti dokument uložený v archivu. V našem případě je u obou uvedena sigla[109] Národní knihovny ČR, která je v roli tvůrce i správce dat. Veškeré informace, které jsou do metadat ukládána, jsou normalizovaná – buď podléhají nějakému kontrolovanému slovníku nebo nějakému standardu, např. datum je generováno podle normy ISO 8601.

```
<mets:metsHdr CREATEDATE="2013-10-25T13:22:13Z"
LASTMODDATE="2013-10-25T13:22:13Z">
      <mets:agent ROLE="CREATOR" TYPE="ORGANIZATION">
            <mets:name>ABA001</mets:name>
      </mets:agent>

      <mets:agent ROLE="ARCHIVIST" TYPE="ORGANIZATION">
            <mets:name>ABA001</mets:name>
      </mets:agent>
</mets:metsHdr>
```

Na dalším příkladě už jsou zobrazeny popisné údaje, které jsou u sklizně. V tomto případě je použitý metadatový standard MODS vložený do hlavního METSu, který ale slouží zejména pro bibliografický popis klasických dokumentů. Pro zachování konzistence je využíván i pro data z Webarchivu. Těchto popisných údajů není velké množství, a proto by bylo kontraproduktivní zavádět další metadatový formát.

```
<mods:mods>
      <mods:titleInfo>
            <mods:title>Serials-2012-01-1M</mods:title>
      </mods:titleInfo>
      <mods:originInfo>
            <mods:dateCaptured encoding="w3cdtf">2012-01-01T13:22:13Z
            </mods:dateCaptured>
      </mods:originInfo>
      <mods:recordInfo>
            <mods:recordIdentifier type="uuid">868fd050-9549-005056829d4
            </mods:recordIdentifier>
      </mods:recordInfo>
      <mods:note type="description">
            sklizen s mesicni a dvou mesicni frekvenci - leden 2012
      </mods:note>
</mods:mods>
```

[109] Sigla je systém jednoznačné identifikace institucí v ČR. Spravuje ji NK ČR.

Příklad ukazuje, že sklizeň obsahuje opravdu málo popisných informací, ty se prakticky skládají pouze z názvu sklizně, data jejího zahájení, poznámek kurátora a přiděleného identifikátoru, o kterém byla zmínka výše. Nicméně již tak málo údajů stačí k úplné identifikaci sklizně, např. podle názvové konvence. V uvedeném příkladu je vidět, že se jedná o výběrovou sklizeň českých internetových periodik, které vycházejí jedenkrát až dvakrát měsíčně, z ledna roku 2012. Pokud by kurátor nebo uživatel chtěl zpětně zkoumat průběh sklizně, musel by si stáhnout celý balíček s logy, reporty a soubory s nastavením. Tyto údaje slouží pouze ke spolehlivému vyhledání sklizně v úložišti dlouhodobé ochrany.

Poslední část metadat u sklizně je tzv. *file section*, která obsahuje seznam všech souborů patřících do balíčku sklizně. Kvůli délce mapy obsahuje příklad pouze první položku v seznamu. O každé položce obsahuje element *file* základní údaje, které jsou uloženy v jeho atributech. Jedná se o vnitřní identifikátor, pořadí (atribut *SEQ*), velikost souboru, datum jeho vytvoření a konečně jeho hash[110], který slouží k ověření, že soubor nebyl změněn nebo poškozen. Tento seznam pro balíček sklizně obsahuje kompletní výčet logů, reportů a souborů s nastavením.

```
<mets:fileSec>
      <mets:fileGrp ID="LOGSGRP" USE="Logs">
            <mets:file CHECKSUM="1fb6d15f19ea79a7c05daa26b6b724b3"
                  CHECKSUMTYPE="MD5"
                  CREATED="2012-01-19T10:14:39"
                  ID="LOG_0001"
                  MIMETYPE="text/plain"
                  SEQ="0"
                  SIZE="13158">
      <mets:FLocat LOCTYPE="URL" xlink:href="LOGS/crawl-manifest.txt"/>
</mets:file>
      ...
</mets:fileSec>
```

Na dalším příkladě jsou popsána metadata, která se nachází v balíčku se samotnými daty. Nejprve je popsán soubor hlavní METS, který je velmi podobný tomu u sklizně. Opět obsahuje metadatový formát MODS zabalený v kontejnerovém formátu METS, a to včetně *file section*, která tentokrát obsahuje seznam dat (jeden ARC balíček) a seznam vyextrahovaných souborů s textem. METS hlavička je úplně stejná jako u sklizně, a tak jsou v příkladu uvedena rovnou popisná metadata. V tomto případě metadata popisují

[110] Hash je výstup hashovací funkce, která slouží k převodu vstupních dat do relativně malého řetězce znaků.

jeden soubor ARC, pro který se používá jako titul název souboru. Stejně jako u sklizně, tak je i u ARCu datum jeho vytvoření a je mu přidělen identifikátor UUID. V elementu *relatedItem* je umístěn identifikátor sklizně, tím je ARC přidělen ke konkrétní sklizni.

```
<mods:mods>
      <mods:titleInfo>
      <mods:title>
SERIALS-2012-01-1M-20120113130837-01118-crawler05.webarchiv.cz.arc.gz
      </mods:title>
      </mods:titleInfo>

      <mods:originInfo>
            <mods:dateCaptured>2012-01-01T14:04:51Z</mods:dateCaptured>
      </mods:originInfo>

      <mods:recordInfo>
            <mods:recordIdentifiertype="uuid">868fd050-3d65-9549-0050568209d4
            </mods:recordIdentifier>
      </mods:recordInfo>

      <mods:relatedItem>
            <mods:recordIdentifier type="uuid">868fd050-9549-0050568209d4
            </mods:identifier>
      </mods:relatedItem>

</mods:mods>
```

5.5.3. Technická a administrativní metadata

Další metadatový soubor obsahuje administrativní a technická metadata. Ta obsahují podrobné informace o souborovém formátu, ve kterém jsou data uložena. Opět je použit strukturální metadatový standard METS, ale tentokrát jsou informace uloženy v metadatovém formátu PREMIS. V tomto souboru je podrobně popsán podle logiky PREMISu objekt (v orig. object) ARC, ke kterému jsou přidruženy události (v orig. events), které byly s objektem prováděny v rámci digitální ochrany a tzv. agenti (v orig. agents), kteří jsou za konkrétní události zodpovědní. Každá činnost, ať minulá nebo budoucí, by měla být právě v souboru s PREMISem uložena. Pokud například dojde v budoucnosti k migraci do nového formátu WARC, pak tato událost musí být zde zaznamenána včetně nástroje, který k tomu byl použit.

Samotný popis objektu je poměrně dlouhý, proto je zde uvedena jako příklad pouze jeho část. V popisu objektu jsou obsaženy identifikátory, velikost souboru nebo třeba hash včetně aplikace, která jej generovala. Také jsou zde pomocí vnitřních identifikátorů nalinkovány události, které se k danému objektu vztahují.

```
<premis:format>
```

```
      <premis:formatDesignation>
            <premis:formatName>application/arc</premis:formatName>
            <premis:formatVersion>1.1</premis:formatVersion>
      </premis:formatDesignation>
      <premis:formatRegistry>
      <premis:formatRegistryName>PRONOM<premis:formatRegistryName>
      <premis:formatRegistryKey>fmt/410</premis:formatRegistryKey>
      </premis:formatRegistry>
</premis:format>
…
<premis:preservationLevel>
      <premis:preservationLevelValue>preservation
      </premis:preservationLevelValue>
      <premis:preservationLevelDateAssigned>2013-10-25
      <premis:preservationLevelDateAssigned>
</premis:preservationLevel>
```

Na příkladu je vidět popis formátového souboru, který obsahuje název formátu včetně jeho verze, a také identifikátor z registru souborových formátů PRONOM[111]. V další části příkladu jsou informace o nastavení úrovně ochrany a datum tohoto nastavení. Zde je konkrétně nastaveno, že daný objekt má být pod ochranou. Ale opět, pokud by hypoteticky v budoucnosti došlo k migraci do formátu WARC, pak by byl starší formát nejspíše zneplatněn a došlo by ke změně úrovně ochrany na smazáno nebo nechráněno.

Na příkladu migrace z ARC do WARC je znázorněno, že metadata nejsou po vygenerování zakonzervována, ale musí v nich být reflektována každá činnost, která je s daty v úložišti prováděna, resp. každá činnost, která do dat nějakým způsobem zasahuje.

[111] PRONOM je webový registr souborový formátů, který vznikl na podporu aktivit spojených s dlouhodobou ochranou digitálních dokumentů

5.6. Správa dat v LTP systému

Jakmile data doputují do dlouhodobého úložiště, tak k nim získají přístup jeho kurátoři. NK ČR používá pro správu dat v dlouhodobém úložišti nástroj zvaný LTP SAFE, který pro ni vyvíjí firma AIP SAFE. Kurátoři mohou data prohlížet, vyhledávat nad nimi, transformovat je a exportovat ze systému.

Vyhledávání probíhá nad metadaty, jejichž extrakce byla popsána v předchozích kapitolách a které byly uloženy do XML souborů. Z těchto souborů si LTP systém vytáhl potřebné elementy, které uložil do své databáze, se kterou lze pracovat v reálném čase, neboť s daty, která jsou archivována na magnetických páskách, toto nelze.

LTP systém také sleduje souborové formáty, ve kterých je obsah uložený, a jeho kurátoři mají nástroje pro jejich monitorování, zda nezastarávají. Export je prováděn do DIP balíčků, které mají strukturu definovanou stejně jako balíčky archivní.

Pokud dojde ke zjištění, že některý ze souborových formátů již nevyhovuje, tak budoucí transformace dat je iniciována z prostředí LTP systému. Ale samotné transformaci musí předcházet analýza nebo vývoj nástrojů, které ji provedou, a tyto nástroje je třeba připojit k transformačnímu modulu. Nástroje musí projít důkladným testováním, protože jakákoliv manipulace s daty je ve všech případech riziková. V LTP systému se data nikdy nepřepisují, ale vytváří se nová verze dokumentu. Kurátoři LTP mají stálý přístup i k minulým verzím dokumentů.

5.7. Další výzkum

V předchozích částech byly popsány nástroje dlouhodobé ochrany a současná situace českého webového archivu. Dnes jsou data z českého webového archivu chráněna pouze na fyzické úrovni pomocí bitové ochrany, která *„se nejčastěji realizuje redundantním uložením a řízeným kopírováním na nové nosiče“*[112]. Tato úroveň ochrany je samozřejmě nedostatečná a český webový archiv pracuje na vytvoření strategie dlouhodobé ochrany svých dat.

Vybudování strategie dlouhodobé ochrany není pro webový obsah lehký úkol. Musejí mu předcházet kroky, které jsou náročné nejen na čas, ale i počítačový výkon. Prvními kroky jsou definování určené skupiny a vytvoření profilu webového archivu.

5.7.1. Určená skupina

Určená skupina (ang. designated community) je termín, který zavádí norma OAIS, je definována jako: *„stanovená skupina možných koncových uživatelů, kteří by měli být schopni porozumět konkrétní množině informací. Určená skupina je vymezena daným archivem a toto vymezení se může časem měnit. Určená skupina se také může skládat z několika uživatelských komunit.“*[113] Definování určené skupiny uživatelů je klíčovým krokem při stanovování strategie dlouhodobé ochrany. Ovlivňuje, jakým způsobem budou data ukládána a jak s nimi bude nakládáno v rámci zavádění strategií dlouhodobé ochrany.

Data v archivu musejí být archivována v takové podobě, aby s nimi určená skupina uživatelů uměla pracovat bez pomoci odborníka. V případě webového obsahu zaměřeného na běžné uživatele se očekává, že uživatel webového archivu bude mít možnost prohlížet si archivní kopie webových stránek tak, jak je tomu zvyklý u živého webu. Webový archiv, který cílí na badatele, musí svá data zpřístupnit, aby mohli využít jejich potenciál (celé datasety, nabídnout API apod.).

---

[112] KVAŠOVÁ, Zuzana a Tomáš SVOBODA. Dlouhodobá ochrana elektronických publikací. *ProInflow* [online]. 2013, Vol. 5, No. 2 . Dostupné z: http://www.phil.muni.cz/journals/index.php/proinflow/article/view/775

[113] ČSN ISO 14721. Systémy pro přenos dat a informací z kosmického prostoru – Otevřený archivační informační systém – Referenční model. Praha: Úřad pro technickou normalizaci, metrologii a státní zkušebnictví, Srpen 2014.

U českého webového archivu jsou definovány tyto určené skupiny:[114]

1. Individuální uživatelé
2. Institucionální uživatelé
3. Výzkumníci a vědci

U individuálních uživatelů se počítá se zájmem o běžnou webovou zkušenost, zprostředkovanou připojením k internetu a webovým prohlížečem. „*Z hlediska požadavku na srozumitelnost informací uložených v archivu je cílem WebArchivu uchovávat a zpřístupňovat uchované informace v takové formě, aby byla co nejvíce podobná formě informací nacházejících se v daném okamžiku na tzv. živém webu.*“[115] Institucionální uživatele zastupují instituce, které potřebují data z webového archivu ke své práci, nyní se v praxi lze setkat s požadavky od soudů, policie či jiných úřadů. Pro tento typ uživatelů jsou požadavky vyřizovány vždy individuálně, mnohdy jsou předávány webové stránky v analogové formě či přímo zdrojové soubory. Institucionální uživatele vždy zajímá více autenticita obsahu než uživatelský zážitek.

Výzkumníci a vědci potenciálně potřebují ke svému výzkumu velké objemy dat z webového archivu a k nim potřebují přímý přístup nebo nástroje, které jim ho zajistí, nebo umožní nad nimi provádět rozsáhlé analýzy. Český webový archiv v současné době pracuje na tom, aby mohl zpřístupnit svá data pro badatelské účely, ať už ve formě celých datových setů nebo pomocí specializovaných nástrojů.

Definování určené skupiny uživatelů webového archivu je jen jeden z prvních kroků pro úspěšné zavedení strategie dlouhodobé ochrany. Dalším krokem je zjištění, jaký obsah je uložený v kontejnerových formátech, resp. v archivu.

5.7.2. Profil webového archivu a formátová analýza obsahu

Vytvoření profilu webového archivu je následujícím krokem k vytvoření strategie dlouhodobé ochrany. Profil kolekce je důležitý zejména pro nehomogenní typ dat, jako jsou právě ta z webového archivu. Pro webové archivy obecně platí, že jejich obsah je

---

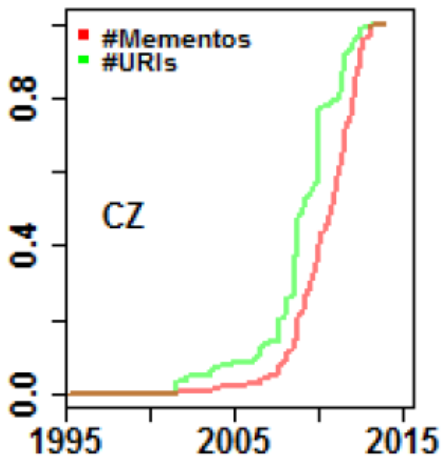
[114] KVASNICA, Jaroslav a Barbora BJAČKOVÁ. Profilování kolekce a stanovení určené skupiny WebArchivu Národní knihovny ČR. ProInflow [online]. 2014, **6**(No. 1) [cit. 2015-09-02]. ISSN 1804–2406. Dostupné z: http://www.phil.muni.cz/journals/index.php/proinflow/article/view/942

[115] KVASNICA, Jaroslav a Barbora BJAČKOVÁ. Profilování kolekce a stanovení určené skupiny WebArchivu Národní knihovny ČR. ProInflow [online]. 2014, **6**(No. 1) [cit. 2015-09-02]. ISSN 1804–2406. Dostupné z: http://www.phil.muni.cz/journals/index.php/proinflow/article/view/942

determinován svým původem, objevuje se v nich nezměrné množství souborových formátů, data jsou mezi sebou nahodile propojena pomocí hypertextových odkazů. Profil archivu je charakteristika archivu jako celku a tvoří jej *„sada charakteristik, která popisuje obsah archivu. Profil je obecný popis na nejvyšší úrovni. Tento popis shrnuje obsah webového archivu.“*[116] Profilace webového archivu pomáhá v rozhodování, na jaké další kroky se zaměřit, odhalit případná rizika, a umožňuje se kvalifikovaně rozhodovat při plánování dlouhodobé ochrany obsahu.

Mezi nejzákladnější charakteristiky patří stáří archivu, tempo nárůstu objemu dat, složení TLD[117] domén nebo poměr souborových formátů v archivu. Český webový archiv byl součástí výzkumu Standfordské univerzity, která profilovala několik webových archivů po celém světě a vzájemně je srovnávala. Z těchto dat a vlastních měření Webarchiv připravil návrh profilu českého webového archivu (Příloha 3).

Vznik českého webového archivu se datuje od roku 2001, stáří archivů je určováno pomocí data nejstarší archivní kopie. Tempo nárůstu objemu dat je exponenciální, jak je vidět na přiloženém grafu (obrázek č. 5: tempo nárůstu objemu dat[118]).



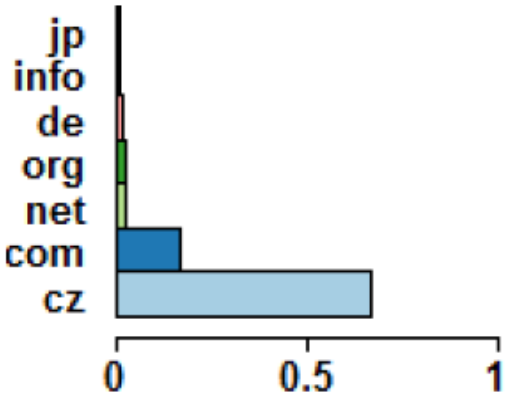
Obrázek č. 5: tempo nárůstu objemu dat

[116] ALSUM, Ahmed, Michele C. WEIGLE, Michael L. NELSON a Herbert VAN DE SOMPEL. Profiling web archive coverage for top-level domain and content language. International Journal on Digital Libraries [online]. 2014 [cit. 2015-07-15]. DOI: 10.1007/s00799-014-0118-y. Dostupný z: http://link.springer.com/10.1007/s00799-014-0118-y

[117] TLD (Top level domain) je zkratka pro doménu prvního řádu, např. .cz, .com.

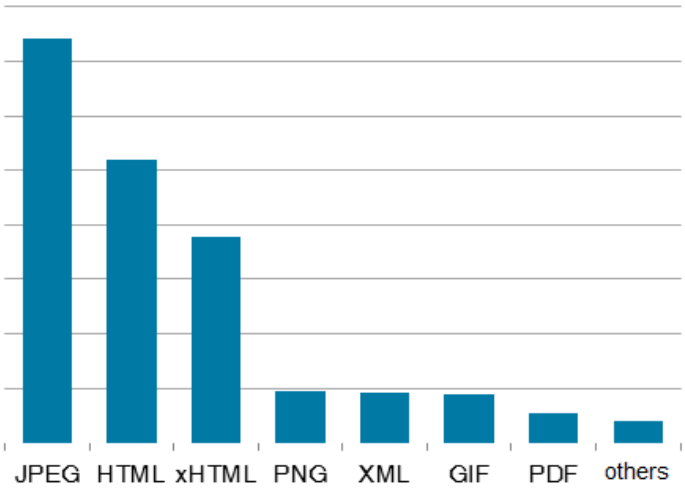
[118] Tamtéž jako 116.

Archivu jednoznačně dominuje česká doména, což koresponduje se zaměřením Webarchivu (obrázek č. 6: rozložení TLD domén).



Obrázek č. 6: rozložení TLD domén

Při charakterizaci souborových formátů, která byla provedena na vzorku dat, se ukázalo, že cca 80 % archivu tvoří 7 souborových formátů, a zbytek je tzv. long tail a není zřejmé, jaký počet souborových formátů může obsahovat. Odhadem to může být v řádech stovek až tisíců (obrázek č. 7: poměr souborových formátů).



Obrázek č. 7: poměr souborových formátů

Definice určené skupiny uživatelů a vytvoření návrhu profilu kolekce jsou pouze počátečními kroky pro zajištění dlouhodobé ochrany webového archivu a v budoucnosti musí následovat další. V současnosti není stále jasné, o jaké kroky se jedná, ale je nutné je určit na základě rizik, která z dosavadního postupu při budování strategie dlouhodobé ochrany vyplývají.

6. Závěr

Webové archivy po celém světě začínají reagovat na požadavky vědců, kteří potřebují mít takový přístup k jejich datům, aby s nimi mohli pracovat stejně jako s jiným typem velkých dat (big data). Tyto požadavky se ozývají zejména z řad novodobých historiků, lingvistů, politologů nebo sociologů, kteří v dnešní době mají pokročilé znalosti s prací na počítači, a tak volají po zpřístupnění celých datových setů či přístupu k datům přes API. Z těchto důvodů se data z webových archivů přesunují na distribuovaná rozšiřitelná úložiště, která jsou navržena pro práci s velkými daty. Tento nový přístup přináší pro webové archivy nové možnosti i v oblasti dlouhodobé ochrany digitálních dokumentů.

Pozornost webových archivů se nyní upírá na vývoj emulačních frameworků, které se zdají být ideálním řešením pro zpřístupnění těch nejstarších dat, která u některých archivů pocházejí již z konce devadesátých let. Emulační frameworky, které běží na cloudu, přinášejí výhodu v jednodušším nasazení pro uživatele, a přestože se v současné době s nimi pouze experimentuje a nelze je označit za plnohodnotné řešení pro dlouhodobou ochranu dat, tak právě snadná použitelnost přispívá k jejich rozšíření. Dnes zejména ve formě služby pro uživatele, která umí zprostředkovat autentický zážitek z historie internetu.

V České republice pomohl rozvířit diskuzi ohledně dlouhodobé ochrany webového obsahu projekt Národní digitální knihovny, který přinesl NK ČR prostředky na vybudování LTP úložiště a v němž již od začátku bylo počítáno s uložením dat webového archivu. Díky tomu mohly být zahájeny práce na celkové koncepci této problematiky.

Koncepce uložení dat z webového archivu do LTP úložiště vznikala na půdě NK ČR a Webarchiv je jedním z prvních webových archivů na světě, který má možnost ukládat svá data na plnohodnotném LTP úložišti. V rámci konsorcia IIPC probíhají diskuze o strategiích a koncepcích pro dlouhodobou ochranu a například Finská národní knihovna zvolila pro ukládání dat obdobnou strategii jako Webarchiv.

Strategie dvou intelektuálních entit (jedna pro data, druhá pro sklizeň) se časem ukázala jako správná cesta, ale i tak je nutné počítat s určitými riziky. V první řadě se jedná o dlouhodobou ochranu pouze na úrovni kontejnerů, na níž musí být nazíráno jen jako na dočasné řešení, neboť ignorování obsahu v kontejnerech může sice stačit v následujících letech, ale jakmile se začnou objevovat indicie, že některý z dominantních

formátů v archivu (80 % webového archivu tvoří pouze 7 souborových formátů) začíná zastarávat, tak začne být situace kritická.

Příští kroky českého webového archivu v oblasti dlouhodobé ochrany dat by měly vést ke kompletní charakterizaci všech souborových formátů, které se v archivu nacházejí. To umožní pracovat z daty, která jsou v kontejnerech. Bohužel dnes tomu brání nedostatečný výpočetní výkon, a to jak na straně samotného webového archivu, kde je nutný přechod na nový typ infrastuktury, tak na straně LTP úložiště, kdy proces ingestu webového obsahu neúměrně zatěžuje celý systém.

Další riziko v současné strategii českého webového archivu je rozhodnutí o migraci staršího kontejnerového formátu ARC do novějšího formátu WARC. Tato migrace má svá opodstatnění, ale přináší s sebou velké riziko poškození dat. K tomuto kroku se zatím odhodlalo jen několik málo archivů ve světě, takže na ověření nástrojů v praxi na větším objemu dat se stále čeká. Webarchiv proto po migraci ARC soubory nemaže, ale ukládá je spolu s novými soubory do AIP balíčku. Ale i přesto by případná chyba při migraci přinesla enormní náklady navíc.

Stejně jako všechny webové archivy a LTP úložiště musí čelit i ty české rizikům plynoucím z ekonomické a personální situace. Bez spolehlivého financování a dostatečného personálního zabezpečení není možné na dlouhodobou ochranu dat pomýšlet. Všechny podobně zaměřené archivy musí mít jasně nastavenou politiku financování na dlouhou dobu dopředu a dostatečnou zastupitelnost v řadách odborníků spravujících archiv, aby se nemohlo stát, že dojde k paralyzování archivu z důvodu nedostatku finančních prostředků nebo odborníků. Dnes toto bohužel v českém prostředí chybí.

Nejdůležitějším přínosem práce je kompletní analýza procesu uložení dat z webového archivu do LTP úložiště. V ní je popsána cesta, kterou putují data v rámci celého systému NDK. Dále přináší nejen popis metadatové strutkury, ale také vzniku a zdrojů metadat. Vysvětlení, jak celá koncepce vznikala, včetně odůvodnění jednotlivých kroků, pomůže případným budoucím odborníkům navázat na dosavadní stav a pokračovat v dalším výzkumu dlouhodobé ochrany webového obsahu.

Zejména výzkum strategií, jako je migrace nebo emulace, je zvláště potřebný a tato diplomová práce může pro něj sloužit jako teoretický úvod. Nicméně vývoj nebo

implementace emulačních frameworků do českého prostředí, či kompletní charakterizace souborových formátů se ukazuje jako úkol, který přesahuje kompetence knihovníků nebo informačních pracovníků a je nutné zapojit do celého procesu více IT odborníků, případně univerzit takového zaměření.

Dlouhodobá ochrana webového obsahu je velmi důležitý úkol, protože se ukazuje, že bez webových archivů se již neobejdou badatelé zabývající se novodobou historií, postupně bude tato potřeba narůstat a zároveň stále více dat bude z živého internetu mizet. Vzhledem k tomu, že je NK ČR v České republice jedinou institucí, která se archivací českého webu zabývá, má velkou zodpovědnost, a tento nelehký úkol bez stabilního financování a soustavného vzdělávání nových odborníků a zapojení dalších institucí do výzkumu nemůže sama zvládnout.

Seznam obrázků a tabulek

Seznam použité literatury

About IIPC. *IIPC: International Internet Preservation Consortium* [online]. 2012. Dostupné z: http://netpreserve.org/about-us

BERMÈS, Emmanuelle a Gautier POUPEAU. Semantic Web technologies for digital preservation: the SPAR project. In: *Proceedings of the Poster and Demonstration Session at the 7th International Semantic Web Conference* [online]. 2008. Dostupné z: http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-401/iswc2008pd_submission_14.pdf

BRENDA AYALA, Reyes. Web Archiving Bibliography 2013. In: *UNT Digital Library* [online]. 28. 6. 2013. Dostupné z: http://digital.library.unt.edu/ark:/67531/metadc172362/m1/1/

BROKEŠ, Adam. Projekt WebArchiv: archiv českého webu. *Zpravodaj ÚVT MU: bulletin pro zájemce o výpočetní techniku na Masarykově univerzitě* [online]. Brno: Ústav výpočení techniky MU, 2008, XVIII, č. 4, s. 10-13. Dostupné z: http://www.ics.muni.cz/zpravodaj/articles/578.html

BROWN, Adrian. Digital Preservation Guidance Note: Selecting Storage Media for Long-Term Preservation. THE UK GOVERNMENT'S OFFICIAL ARCHIVE. *The National Archives* [online]. 2008. Dostupné z: http://www.nationalarchives.gov.uk/documents/selecting-storage-media.pdf

CAPLAN, Priscilla. *Understanding PREMIS* [online]. U.S.A.: The Library of Congress, 2009. Dostupné z: http://www.loc.gov/standards/premis/understanding-premis.pdf

CELBOVÁ, Ludmila. *Archivace webu*. 1. vyd. Praha: Národní knihovna ČR, 2008, 45 s. ISBN 978-80-7050-562-5.

CONWAY, Paul. Preservation in the Age of Google: Digitization, Digital Preservation, and Dilemmas. *Library Quarterly* [online]. Chicago: University of Chicago, 2010, vol. 80, no. 1, s. 61-79. Dostupné z: http://deepblue.lib.umich.edu/bitstream/handle/2027.42/85223/J15%20Conway%20Preservation%20Age%20of%20Google%202010.pdf

COUFAL, Libor. Web po 20 letech: co z něj zbude pro budoucí generace?. *Knihovna* [online]. 2009, roč. 20, č. 2, s. 17-32. Dostupné z: http://knihovna.nkp.cz/knihovna92/0902097.htm

CUBR, Ladislav. *Dlouhodobá ochrana digitálních dokumentů*. 1. vyd. Praha: Národní knihovna České republiky, 2010, 154 s. ISBN 978-80-7050-588-5.

CUBR, Ladislav. *Zpráva ze služební cesty: Návštěva Národní knihovny Francie (BNF) [online]*. Praha, 2012. Dostupné z: https://www.nkp.cz/soubory/ostatni/cz_pariz2012_lc.pdf

Cyberspace, the old-fashioned way. *Rhizome* [online]. NY, USA: New Museum, 2015. Dostupné z: http://rhizome.org/editorial/2015/nov/30/oldweb-today/

ČSN ISO 14721. Systémy pro přenos dat a informací z kosmického prostoru - Otevřený archivační informační systém - Referenční model. Praha: Úřad pro technickou normalizaci, metrologii a státní zkušebnictví, 2014.

DAY, Michael. *Collecting and preserving the World Wide Web* [online]. 2003. Dostupné z:http://www.jisc.ac.uk/uploaded_documents/archiving_feasibility.pdf

DAY, Michael. The Long-Term Preservation of Web Content. MASANÈS, Julien. *Web archiving*. New York: Springer, c2006, s. 177-199. ISBN 3540233385-.

DE SOMPEL, Van. HTTP Framework for Time-Based Access to Resource States: Memento. IETF Tools: ETF-related tools, standalone or hosted on tools.ietf.org. [online]. USA, 2013. Dostupné z: https://tools.ietf.org/html/rfc7089#section-1.1

Definitions of Digital Preservation. ALCTS: Association for Library Collections & Technical Services [online]. Washington, D.C., 2007. Dostupné z: http://www.ala.org/alcts/resources/preserv/defdigpres0408

DVOŘÁK, Tomáš. Uchovávání digitálních dokumentů se zachováním jejich autenticity. *Ústav informačních studií a knihovnictví: Jinonické informační pondělky* [online]. 2010. Dostupné z: http://uisk.ff.cuni.cz/dwn/1003/14262cs_CZ_jip.pdf

Emulation as a Tool for Web Preservation: Authentic Access and Efficient Web-server Preservation. *International Internet Preservation Consortium* [online]. 2016. Dostupné z: http://www.netpreserve.org/sites/default/files/WAC-Thomas_Liebetraut.pdf

ENDERS, Markus. A METS based information package for long term preservation of web archives. In: RAUBER.., Andreas.. a Andreas RAUBER. *IPRES 2010 proceedings of the 7th International Conference on Preservation of Digital Objects ; September 19 - 24, 2010, Vienna, Austria*. Wien: Österreich. Computer Gesellsch, 2010, s. 31-40. ISBN 9783854032625. Dostupné z: http://publik.tuwien.ac.at/files/PubDat_191968.pdf

FARRELL, Susan a Kevin ASHLEY. *A guide to Web preservation: practical advice for web and records managers based on best practices from the JISC-funded PoWR project* [online]. S.l.: UKOLN / ULCC, 2010. ISBN 09-516-8567-8. Dostupné z: http://jiscpowr.jiscinvolve.org/wp/files/2010/06/Guide-2010-final.pdf

GOETHALS, Andrea, Clément OURY, David PEARSON, Barbara SIERMAN a Tobias STEINKE. Facing the Challenge of Web Archives Preservation Collaboratively: The Role and Work of the IIPC Preservation Working Group. D-Lib Magazine [online]. 2015, **21**(5/6), - DOI: 10.1045/may2015-goethals. ISSN 1082-9873. Dostupné z: http://www.dlib.org/dlib/may15/goethals/05goethals.html

GRUBER, Lukáš, Tomáš SÍBEK a Libor COUFAL. Archivace webových stránek v českém prostředí aneb Jak funguje WebArchiv. Čtenář [online]. Kladno: Středočeská vědecká knihovna v Kladně, 2009, **2009, 61**(5/2009). ISSN 1805-4064. Dostupné z: http://ctenar.svkkl.cz/clanky/2009-roc-61/05-2009/tema-archivace-webovych-stranek-v-ceskem-prostredi-aneb-jak-funguje-webarchiv-58-393.htm

HLOUŠEK, Petr. *Problematika dlouhodobého uchovávání digitálních dat*. Brno, 2008. Dostupné z:http://is.muni.cz/th/179500/ff_b/BakalarskaDP.pdf. Bakalářská práce. Masarykova univerztita, Filozofická fakulta, Ústav české literatury a knihovnictví,, Kabinet informačních studií a knihovnictví.

HOEVEN, Jeffrey Van der, Bram LOHMAN a Remco VERDEGEM. Emulation for Digital Preservation in Practice: The Results. *International Journal of Digital Curation* [online]. Bath: UKOLN, University of Bath, 2007, vol. 2, issue 2, s. 207-219. DOI: 10.1007/978-3-540-33640-2_10. Dostupné z: http://50.17.193.184/omeka/files/original/84cca606bbb8f1955f42b22c29268811.pdf

HUTAŘ, Jan, Marek MELICHAR a Bohdana STOKLASOVÁ. Národní digitální knihovna. *Knihovna*. 2009, roč. 20, č. 1. Dostupné z: http://knihovna.nkp.cz/knihovna91/humesto.htm

HUTAŘ, Jan. Podrobnější popis projektu NDK a jeho kontext. NÁRODNÍ KNIHOVNA ČESKÉ REPUBLIKY. *Národní digitální knihovna* [online]. 2010. Dostupné z: https://web.archive.org/web/20110106000746/http://ndk.cz/narodni-dk/podrobnejsi-popis-projektu/podrobnejsi-popis-projektu-ndk

ISO 28500:2009. *WARC file format*. 1. vyd. Londýn: British Standard Institute, 2009

JACKSON, Andy. User driven digital preservation with Interject. THE BRITISH LIBRARY. *UK Web Archive blog* [online]. 2014. Dostupné z: http://britishlibrary.typepad.co.uk/webarchive/2014/08/user-driven-digital-preservation-with-interject.html

KLEIN, Lauren. An Archive of Tweets? In: Media, Materiality, and Archives [online]. Atlanta, Georgia, 2016. Dostupné z: http://blogs.iac.gatech.edu/archives16/2016/01/26/an-archive-of-tweets/

KRATOCHVÍLOVÁ, Zuzana. Dlouhodobá ochrana a zpřístupnění dat z webových archivů: WebArchiv Národní knihovny České republiky. *Knihovna: knihovnická revue*. 2012, roč. 23, č. 2., s. 35-47. Dostupné z:http://knihovna.nkp.cz/knihovna122/kratochv.htm

KVASNICA, Jaroslav a Barbora BJAČKOVÁ. Profilování kolekce a stanovení určené skupiny WebArchivu Národní knihovny ČR. ProInflow [online]. 2014, **6**(No. 1). ISSN 1804–2406. Dostupné z: http://www.phil.muni.cz/journals/index.php/proinflow/article/view/942

KVASNICA, Jaroslav a Rudolf KREIBICH. Formátová analýza sklizených dat v rámci projektu WebArchiv NK ČR. *ProInflow*. 2013, 5. ročník, 2. číslo. Dostupné z: http://pro.inflow.cz/formatova-analyza-sklizenych-dat-v-ramci-projektu-webarchiv-nk-cr

KVASNICA, Jaroslav a Rudolf KREIBICH. Specifikace pro data z WA: pouze pro formát WARC. Praha, 2014.

KVAŠOVÁ, Zuzana a Tomáš SVOBODA. Dlouhodobá ochrana elektronických publikací. *ProInflow* [online]. 2013, Vol. 5, No. 2 . Dostupné z: http://www.phil.muni.cz/journals/index.php/proinflow/article/view/775

LAZORCHAK, Butch. Web Archive Preservation Planning. In: *Library of Congress* [online]. 18. 8. 2011. Dostupné z: http://blogs.loc.gov/digitalpreservation/2011/08/web-archive-preservation-planning/

LONG, Andrew Stawowczyk. *Long-term preservation of web archives: Experimenting with emulation and migration methodologies* [online]. Australia: National Library of Australia, 2009. Dostupné z: http://www.netpreserve.org/sites/default/files/resources/Methodologies.pdf

LUKŠŮ, Alžběta. *Dlouhodobé uchovávání a zpřístupňování dokumentů zaznamenaných na optických discích*. Brno, 2010. Dostupné z: http://is.muni.cz/th/262809/ff_b/Bakalarska_prace_Alzbeta_Luksu.txt. Bakalářská práce. Masarykova univerzita.

LUKŠŮ, Alžběta. Dlouhodobé uchovávání digitálních dokumentů. In: *WikiKnihovna: Knihovníci sobě* [online]. 26. 4. 2010, 6. 2. 2012. Dostupné z: http://wiki.knihovna.cz/index.php?title=Dlouhodobé_uchovávání%C3%AD_digitálních%C3%ADch_dokumentů#Pou.C5.BEit.C3.A9_zdroje

MASANÈS, Julien. *Web archiving*. New York: Springer, c2006, vii, 234 p. ISBN 35-402-3338-5.

MELICHAR, Marek a Jan HUTAŘ. České paměťové instituce a digitální data: historický exkurz, současný stav a předpokládaný vývoj III. *Duha* [online]. 2014, roč. 28, č. 2. Dostupné z: http://duha.mzk.cz/clanky/ceske-pametove-instituce-digitalni-data-historicky-exkurz-soucasny-stav-predpokladany-vyvoj-1

OURY, Clément a Sébastien PEYRARD. From the World Wide Web to digital library stacks: preserving the French web archives. In: Proceedings of the 8th International Conference on Preservation of Digital Objects (iPRES) [online]. Singapore: National Library Board : Nanyang Technological University, 2011, s. 237-241. ISBN 978-981-07-0441-4. Dostupné z: https://halshs.archives-ouvertes.fr/halshs-00868729/document

PENNOCK, Maureen. Web-Archiving. *DPC Technology Watch Series* [online]. 2013, roč. 13, č. 01. DOI: http://dx.doi.org/10.7207/twr13-01. Dostupné z: http://dx.doi.org/10.7207/twr13-01

Podrobnější popis projektu NDK a jeho kontext. Portál Národní digitální knihovny [online]. Praha, 2011. Dostupné z: http://wayback.webarchiv.cz/wayback/20140320103556/http://ndk.cz/narodni-dk/podrobnejsi-popis-projektu

Preservation Is Knowledge: A community-driven preservation approach. In: *IPRESS 2012: Proceedings of the 9th International Conference on Preservation of Digital Objects* [online]. Toronto, 2012. Dostupné z: http://www.bnf.fr/documents/ipress2012_art_spar.pdf

Preservation policy. NATIONAL LIBRARY OF AUSTRALIA. *National Library of Australia* [online]. 2009. Dostupné z: http://www.nla.gov.au/policy-and-planning/preservation-policy

ROSENTHAL, David S.H. Emulation & Virtualization as Preservation Strategies. In: The Andrew W. Mellon Foundation [online]. New York: The Andrew W. Mellon Foundation, 2015. Dostupné z: https://mellon.org/Rosenthal-Emulation-2015

Selecting the right preservation strategy. *Paradigm: The Personal Archives Accessible in Digital Media* [online]. 2008. Dostupné z: http://www.paradigm.ac.uk/workbook/preservation-strategies/selecting-emulation.html

STRODL, Stephan, Peter Paul BERAN a Andreas RAUBER. Migrating Content in WARC Files. In: MASANES, Julien a Andreas RAUBER. *The 9th International Web Archiving Workshop (IWAW 2009) Proceedings* [online]. Paris (France): European Archive Foundation, 2009, s. 43-49. Dostupné z: http://publik.tuwien.ac.at/files/PubDat_181115.pdf

SVOBODA, Tomáš. Projekt Národní digitální knihovna: aktuální stav projektu. In: *INFORUM 2012: 18. konference o profesionálních informačních zdrojích*. Praha, 2012, s. 1-5. Dostupné z: http://www.inforum.cz/pdf/2012/svoboda-tomas.pdf

ŠVÁSTOVÁ, Pavla a Jaroslav KVASNICA. Definice metadatových formátů pro digitalizaci monografických dokumentů (monografií, kartografických dokumentů, hudebnin). NÁRODNÍ KNIHOVNA ČESKÉ REPUBLIKY. Národní digitální knihovna [online]. 2013. Dostupné z: http://www.ndk.cz/archivace/DMF-monografie-1-1.pdf

Update on the Twitter Archive At the Library of Congress. Library of Congress [online]. Washington, D.C.: Library of Congress, 2013. Dostupné z: https://www.loc.gov/today/pr/2013/files/twitter_report_2013jan.pdf

VOJTÁŠEK, Filip. Dlouhodobá archivace digitálních dokumentů. *Ikaros* [online]. 2000, roč. 4, č. 10. Dostupné z: http://www.ikaros.cz/dlouhodoba-archivace-digitalnich-dokumentu

Web Archiving Guidance. *The National Archives: The UK government's official archive* [online]. 2011. Dostupné z: http://www.nationalarchives.gov.uk/documents/information-management/web-archiving-guidance.pdf

Web Archiving: Issues and Methods. MASANÈS, Julien. *Web archiving*. New York: Springer, c2006, s. 2-53. ISBN 3540233385-.

WebArchiv: získávání, archivace a zpřístupnění domácích webových zdrojů. *Ikaros* [online]. 2004, roč. 8, č. 5/2. Dostupné z: http://www.ikaros.cz/node/1638

ZACH, Michael. *Celosvětový Archiv Internetu a jeho role v získávání, uchování a zpřístupňování webových zdrojů*. Praha, 2007. Bakalářská práce. Univerzita Karlova v Praze. Vedoucí práce PhDr. Eva Bratková.

Příloha 1: Specifikace pro data z WA (ARC)

Specifikace pro data z WA (pouze pro formát ARC)

29. 1. 2014

Za NK ČR sepsali Jaroslav Kvasnica, Rudolf Kreibich

Obsah

1. Adresářová struktura

/
*.arc.gz
*.arc.gz.open -- nahrát stejně jako formát arc.gz

./logs
*harvest.xml [mandatory]
./dmdsec [mandatory]
*.xml

./index [mandatory]
*.cdx

./crawl [optional] [tar.gz]
./logs
*.log
./settings
./<domena>
./<subdomena>
./..
*.xml
order*.xml
seeds.txt
state.job
seeds-report.txt
responsecode-report.txt
processors-report.txt
mimetype-report.txt
crawl-manifest.txt
crawl-report.txt
frontier-report.txt
hosts-report.txt
.

Příklad:

./Serials-2012-01-1M
SERIALS-2012-01-1M-20120112075013-00000-crawler05.webarchiv.cz.arc.gz
…
./logs
Serials-2012-01-1Mharvest.xml

./crawl // může obsahovat více archivů
Serials-2012-01-1M-00.tar.gz
Serials-2012-01-1M-01.tar.gz
…

./dmdsec
Mets_abclinuxu.cz.xml
…
./index
Serials-2012-01-1M.cdx

2. Struktura balíčků SIP

Balíček pro ARC

SLOŽKA>	OBSAHUJE>>	OBSAHUJE>>>
ARC/WARC	info.xml	
	data(složka)	arcs soubory
	TXT (složka)	txt soubory
	amdSec (složka)	metadata
	hlavní_METS.xml	
	soubor.md5	

Balíček pro sklizeň

SLOŽKA>	OBSAHUJE>>	OBSAHUJE>>>
Sklizeň	info.xml	
	data(složka)	soubory s logy, reporty a nastavením = celý adresář ./logs
	hlavní_METS.xml	
	soubor.md5	

Výstup: SIP = DIP

3. Metadata

Metadata pro balíček s ARC

Hlavní METS:
- příloha: METS_ed5d9aa0-3d66-11e3-9549-0050568209d4-ARC-example.xml

Vedlejší METS:
- příloha: AMD_METS_SERIALS-2012-01-1M-20120113130837-01118-crawler05-ARC-example.xml

Metadata pro balíček se sklizní

Hlavní METS:
- příloha: METS_868fd050-3d65-11e3-9549-0050568209d4-HARVEST-example.xml

4. Doporučení

Identifikátor sklizně: UUID: Generuje naše aplikace Harvester a ukládá do ./logs/ *harvest.xml a do naší MySQL databáze, kde si kontroluje unikátnost. Identifikátor je uložený v Dublin Core v elementu <dc.identifier>.

Soubor s těmito metadaty je povinný a bude se nacházet u každé sklizně.

Příklad:

```
<mets:mets xmlns:mets="http://www.loc.gov/METS/" xmlns:xlink="http://
www.w3.org/1999/xlink" xmlns:xsi="http://www.w3.org/2001/XMLSchema-
instance" xsi:schemaLocation="http://www.w3.org/2001/XMLSchema-instance
http://www.w3.org/2001/XMLSchema.xsd http://www.loc.gov/METS/ http://
www.loc.gov/standards/mets/mets.xsd">
  <mets:metsHdr ROLE="CREATOR" TYPE="ORGANIZATION">
    <mets:name>ABA001</mets:name>
  </mets:metsHdr>
  <mets:dmdSec ID="DCMD_CRAWL_0001">
    <mets:mdWrap DMTYPE="DC" MIMETYPE="text/xml">
      <mets:xmlData xmlns:dc="http://purl.org/dc/elements/1.1/">
        <oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/
oai_dc/">
          <dc:title>Serials-2012-01-1M</dc:title>
          <dc:type>Crawl</dc:type>
          <dc:audience>ABA 001</dc:audience>
          <dc:identifier>1wdutnp2-mbhs-v77f-qfex-9rczq2o8y794</
dc:identifier>
        </oai_dc:dc>
      </mets:xmlData>
    </mets:mdWrap>
  </mets:dmdSec>
  <mets:fileSec>
   ...
```

Příloha 2: Specifikace pro data z WA (WARC)

Specifikace pro data z WA (pouze pro formát WARC)

27. 1. 2014

Za NK ČR sepsali Jaroslav Kvasnica, Rudolf Kreibich

Obsah

1. Adresářová struktura

/
*.warc.gz
*.warc.gz.open -- nahrát stejně jako formát warc.gz

./logs
*harvest.xml [mandatory]
./dmdsec [mandatory]
*.xml

./index [mandatory]
*.cdx

./crawl [mandatory]
*.tar.gz

Příklad:

./Serials-2012-09-1M_6M/

Serials-2012-09-1M_6M-20120925140111676-00001-30141~crawler01.webarchiv.cz~7778.warc.gz

Serials-2012-09-1M_6M-20120925140209038-00001-30261~crawler00.webarchiv.cz~7778.warc.gz

Serials-2012-09-1M_6M-20121009033353221-00367-30261~crawler00.webarchiv.cz~7778.warc.gz.open

…

./logs

Serials-2012-09-1M_6Mharvest.xml

./crawl

Serials-2012-08-1M_6M-crawler00.tar.gz

./dmdsec

Mets_abclinuxu.cz.xml

…

./index

Serials-2012-09-1M_6M.cdx

2. Struktura balíčků SIP

Balíček pro ARC

SLOŽKA>	OBSAHUJE>>	OBSAHUJE>>>
ARC/WARC	info.xml	
	data(složka)	warcs soubory
	TXT (složka)	txt soubory
	amdSec (složka)	metadata
	hlavní_METS.xml	
	soubor.md5	

Balíček pro sklizeň

SLOŽKA>	OBSAHUJE>>	OBSAHUJE>>>
Sklizeň	info.xml	
	data(složka)	soubory s logy, reporty a nastavením = celý adresář ./logs
	hlavní_METS.xml	
	soubor.md5	

Výstup: SIP = DIP

3. Metadata

Metadata pro balíček s WARC

Hlavní METS:
- příloha: METS_ed5d9aa0-3d66-11e3-9549-0050568209d4-WARC-example.xml

Vedlejší METS:
- příloha: AMD_METS_SERIALS-2012-01-1M-20120113130837-01118-crawler05-WARC-example.xml

Metadata pro balíček se sklizní

Hlavní METS:
- příloha: METS_868fd050-3d65-11e3-9549-0050568209d4-HARVEST-example.xml

4. Doporučení

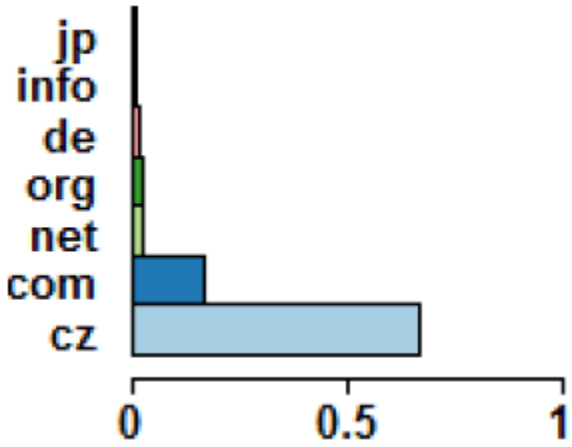
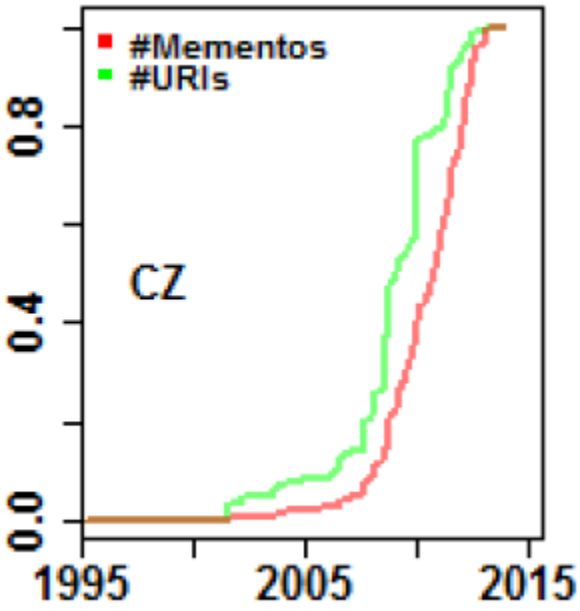
Identifikátor sklizně: UUID: Generuje naše aplikace Harvester a ukládá do ./logs/ *harvest.xml a do naší MySQL databáze, kde si kontroluje unikátnost. Identifikátor je uložený v Dublin Core v elementu <dc.identifier>.

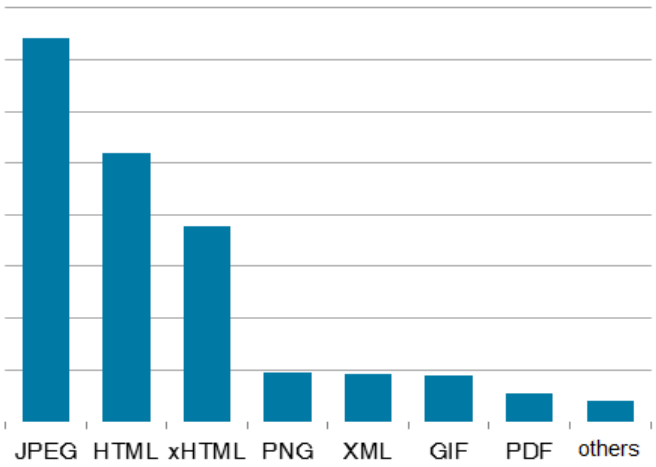
Soubor s těmito metadaty je povinný a bude se nacházet u každé sklizně.

Příklad:

```
<mets:mets xmlns:mets="http://www.loc.gov/METS/" xmlns:xlink="http://
www.w3.org/1999/xlink" xmlns:xsi="http://www.w3.org/2001/XMLSchema-
instance" xsi:schemaLocation="http://www.w3.org/2001/XMLSchema-instance
http://www.w3.org/2001/XMLSchema.xsd http://www.loc.gov/METS/ http://
www.loc.gov/standards/mets/mets.xsd">
  <mets:metsHdr ROLE="CREATOR" TYPE="ORGANIZATION">
    <mets:name>ABA001</mets:name>
  </mets:metsHdr>
  <mets:dmdSec ID="DCMD_CRAWL_0001">
    <mets:mdWrap DMTYPE="DC" MIMETYPE="text/xml">
      <mets:xmlData xmlns:dc="http://purl.org/dc/elements/1.1/">
        <oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/
oai_dc/">
          <dc:title>Serials-2012-01-1M</dc:title>
          <dc:type>Crawl</dc:type>
          <dc:audience>ABA 001</dc:audience>
          <dc:identifier>1wdutnp2-mbhs-v77f-qfex-9rczq2o8y794</
dc:identifier>
        </oai_dc:dc>
      </mets:xmlData>
    </mets:mdWrap>
  </mets:dmdSec>
  <mets:fileSec>
   ...
```

Příloha 3: Návrh profilu českého webového archivu

Profil Webarchivu	
Stáří archivu	3. 9. 2001
Top-level domény	 Obr. 3 Top-level domény
Tempo růstu	 Obr. 4 Tempo růstu
Frekvence sklízení	comprehensive: 1x/year; selective 1x|2x|6x|12x/year
Hloubka sklízení	comprehensive: 5000 objects; selective: 10000-15000 objects
Přístupnost	comprehensive: in house; selective more than 4000 URLs online access
Software	Heritrix Engine 3.1.2-SNAPSHOT-20130207.001528 Wayback-1.5.3-SNAPSHOT

Formáty	 Obr. 5 Přehled formátů
Robot.txt	Don’t respect