Nové standardy digitálních knihoven pro dlouhodobou ochranu

Poznámky, text a překlad
Martin Vojnar
Vědecká knihovna v Olomouci
vojnar@vkol.cz

V průběhu letošního IFLA kongresu (14.–18. srpna 2005 v Oslo) zaznělo několik vystoupení na téma digitálních knihoven. Nejvíce mě z nich zaujal příspěvek Sally H. McCallum (*Library of Congress*) s názvem „Standardizace metadat pro dlouhodobou ochranu elektronických zdrojů: co máme a co potřebujeme“ (*Preservation Metadata Standards for Digital Resources: What we have and What we need*, plný text viz. http://www.ifla.org/IV/ifla71/papers/060e-McCallum.pdf). Charakterizuje současný směr digitálních knihoven, které se přesouvají z výzkumné a vývojové oblasti do provozního režimu, a s tím úzce souvisí i otázka dlouhodobého zachování jejich obsahu. V první části tohoto článku vám proto nabídnu jeho volný překlad. Je to relevantní téma nejen pro správce digitálních a digitalizovaných dat, ale také pro vedoucí projektů digitálních knihoven, neboť s sebou přináší dlouhodobý úkol a výzvu, jaká nemá v historii aplikace informačních technologií v knihovnách srovnatelného předchůdce. V druhé části se pak zaměřím na problematiku výměnného formátu pro prostředí digitálních knihoven. Její spojitost s dlouhodobou ochranou je více než zřejmá a měla by se stát předmětem pokračujících analýz, minimálně v projektech garantovaných na národní úrovni.

Standardizace metadat pro dlouhodobou ochranu elektronických zdrojů: co máme a co potřebujeme

A. Základ metadat pro dlouhodobou ochranu: PREMIS

Počátky

Projekt PREMIS (Strategie implementace metadat pro dlouhodobou ochranu, *The Preservation Metadata Implementation Strategies*, viz. http://www.loc.gov/standards/premis) vznikl na základě zkušeností z posledních deseti let. V knihovním prostředí bylo věnováno hodně práce datovým uložištím, zejména mezi klíčovými členy ICABS (*IFLA/CDNL Alliance for Bibliographic Standards*) a jejich spolupracovníky. Část tohoto úsilí představoval návrh pracující s formálními i neformálními datovými modely, který se snažil nalézt patřičné datové prvky pro zajištění funkce dlouhodobé ochrany, i když jeho širší záběr přesahoval pouhý rámec ochrany a mířil také ke zpřístupnění a šíření. Z příkladů takových projektů lze zmínit projekt NEDLIB (*Networked European Deposit Library*) holandské a francouzské národní knihovny, projekt CEDARS (*CURL Exemplars in Digital Libraries*) z Velké Británie nebo projekt Pandora australské národní knihovny a řadu dalších institucionálních aktivit (*OCLC, Library of Congress* aj.).

Je zajímavé, že všechny uvedené projekty se určitým způsobem dotýkaly referenčního modelu OAIS (*Open Archival Information System*, viz. http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf), který byl primárně vytvořen pro potřebu

datových úložišť a později se stal ISO standardem (ISO 14721). OAIS model tak působil jako jednotící prvek výzkumných činností tím, že poskytl společný jazyk. Díky tomu jsou dnes informační soubory určené pro archivaci (AIP, *Archival Information Package*), dodávání (SIP, *Submission Information Package*) a šíření (DIP, *Distribution Information Package*) všeobecně uznávány coby komponenty základního konceptu implementace digitálních repozitářů. Informační soubory se skládají ze čtyř částí: vlastní obsah, kontejner, popis a dlouhodobá ochrana. V roce 2002 odvedly vynikající práci OCLC a RLG, které integrovaly výsledky výše zmíněných projektů a zasadily je do rámce OAIS modelu (http://www.oclc.org/research/projects/pmwg/pm_framework.pdf). Hlavním cílem pracovní skupiny PREMIS pak bylo na těchto základech vytvořit datový slovník určený k implementaci do systémů.

Cíle

Projekt PREMIS představoval víceleté úsilí pracovní skupiny reprezentované zástupci z Austrálie, Nového Zélandu, Spojených států, Velké Británie, Německa a Holandska. Původně měla práce skupiny trvat jeden rok, ale byla prodloužena na dvojnásobek. Jejím výsledkem byla velmi podrobně strukturovaná sada datových prvků určených k implementaci. Skupina si stanovila během své činnosti několik samostatných cílů, z nichž měly všechny stejný praktický základ a podstatný směr vedoucí k reálné implementaci. Mezi tyto cíle – oba se podařilo splnit – patřila identifikace základního metadatového formátu a vytvoření jeho datového slovníku. Třetím úkolem byla definice možných strategií a doporučení pro implementaci. Ukázalo se, že nejlepším způsobem jejich realizace bylo experimentování s vytvářeným datovým slovníkem. Pokračování prací by mělo dále zahrnovat pilotní testování datového slovníku a kooperace činností založených na základním metadatovém formátu.

Přehled

Nejprve byl sestaven přehled stávajících implementací datových digitálních úložišť, který měl shrnout současné přístupy a trendy. Zúčastnilo se ho 48 respondentů ze 13 zemí, což poskytlo solidní základ. Bylo z něj možné vyvodit následující obecné závěry (http://www.oclc.org/research/projects/pmwg/surveyreport.pdf), které usměrňovaly průběžně pokračující práce na datovém modelu:

- obecně rozšířené použití referenčního modelu OAIS představuje dobrý začátek pro návrh datového modelu úložiště
- většinou datová úložiště ukládají i metadatové informace dvojím způsobem – jednak v XML podobě, jednak v relační databázi pro rychlý přístup a práci s nimi a počítají se specifickým vymezením pro dlouhodobou ochranu
- velmi často je pro kontextové uložení metadat použit formát METS (*Metadata Encoding and Transmission Standard*); používaná schémata metadat se různí, lze v něm počítat i s metadaty pro dlouhodobou ochranu včetně technických metadat pro obrazové objekty (*MIX, Metadata for Images in XML*)
- současným trendem je uchovávat digitální objekty v několika verzích (kromě původní jsou to normalizované podoby pro zpřístupňování, podoby před a po migraci, každá s odpovídajícími metadaty)
- i v rámci jedné instituce je obvyklý vícenásobný přístup ke strategii uchovávání s ohledem na výzkumný a vývojový charakter prací

Dále se v přehledu ukázalo, že byla učiněna řada analýz s ohledem na různou povahu jednotlivých druhů objektů (binární data, soubory, sbírky, logické objekty) a že se zpracovává a ukládá

informace o vztazích mezi objekty. Ačkoliv svou formou přehled neaspiroval na konečný souhrn současného stavu, jeho výsledky byly pro práci na datovém slovníku zajímavé a užitečné.

Datový slovník

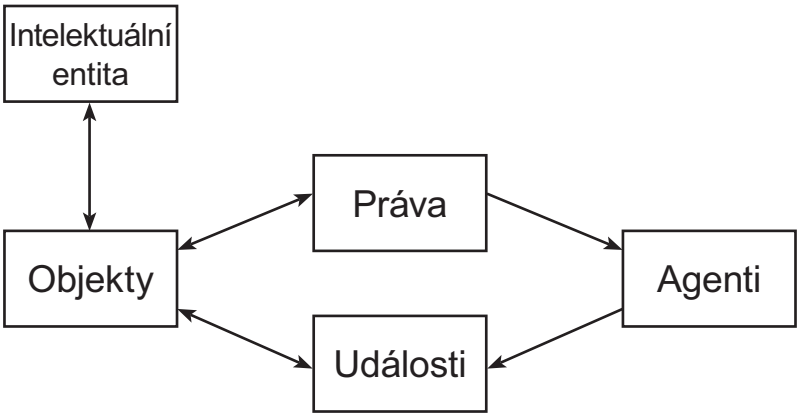
Hlavní prvky datového slovníku (http://www.oclc.org/research/projects/pmwg/premis-final.pdf) byly vyvinuty pracovní skupinou PREMIS na základě přehledu digitálních datových úložišť. Na samotném začátku bylo přijato několik praktických rozhodnutí, která se ukázala jako důležitá. Za jádro datových prvků skupina považovala „údaje, které by chtěla používat většina současných fungujících datových úložišť, aby dosáhla dlouhodobé ochrany". Skupina z plánovaného záběru datového slovníku záměrně vyloučila některé dobře známé aspekty dlouhodobé ochrany, např. detailní technická metadata pro různé druhy objektů. Do návrhu datového slovníku se tak dostala pouze taková technická metadata, která mají pro existující formáty obecný význam.

Další důležitou přijatou zásadou byl fakt, že specifikovaná metadata musí být v maximální možné míře získávána a používána automatizovaným způsobem. Upřednostňována tak byla standardizována metadata z řízených schémat a slovníků oproti volnému textovému popisu, což podpořilo záměr skupiny budovat datový slovník nezávisle na implementaci v konkrétním systému. Jak se totiž ukázalo v přehledu, digitální repozitáře v provozní fázi a ve fázi implementace mohou mít různé vlastnosti s ohledem na své prostředí. Jádro datových prvků PREMIS, ze kterého může datové úložiště čerpat, v něm pak nemusí být nutně také uloženo, ale může být přítomné v nástrojích třetích stran, ať už to bude komerční řešení, se kterým datové úložiště spolupracuje, nebo lokální databáze ve vlastním formátu. Je potřeba zdůraznit předpoklad, že jádro metadat by mělo být k dispozici pro případnou konverzi do určitého standardu v případě změny systému – jinak řečeno, jádro metadat by mělo být k dispozici libovolnému programu, který bude datové úložiště přímo nebo zprostředkovaně používat pro práci s daty uloženými podle datového slovníku PREMIS. Z těchto důvodů se rozhodla pracovní skupina v datovém slovníku definovat tzv. sémantické jednotky namísto metadatových prvků.

Datový model

S ohledem na délku příspěvku zde nebudu uvádět podrobný popis celého datového modelu a zmíním se pouze o některých důležitých rysech (vysvětlení celého modelu lze v úplnosti najít ve zprávě PREMIS). Datový model je jednoduchý. Obsahuje 5 základních typů entit: Objekty (*Objects*), Události (*Events*), Agenty (*Agents*), Práva (*Rights*) a Intelektuální entitu (*Intellectual Entity*) samotnou. S ohledem na jádro datového slovníku byla volba jeho obsahu pečlivě zvažována. Proto byly některý druhy metadat (např. popisných pro popis knihy, mapy, www stránky atd.) ponechány ve prospěch současných standardů (pro popisná metadata např. MARC, MODS, Dublin Core…), které již fungují. Obdobně by měla být ošetřená podrobná informace týkající se agentů, kde lze využít formáty MARC, MADS (*Metadata Authority Description Standard*), vCard a jiné standardy. Údaje týkající se práv byly vymezeny jejich vztahem k činnostem, kterými se provádí dlouhodobá ochrana, zatímco práva pro zpřístupňování a šíření se nepovažují za součást tohoto jádra. Podrobná technická metadata a dokumentace fyzických zařízení (hardware) sice také v modelu zahrnuta nejsou, ale lze je použít podle potřeby a specifikace.

Sémantické jednotky pro objekty (kterým je v modelu věnována zásadní pozornost) mohou být specifikovány ve třech úrovních s cílem umožnit flexibilitu datového úložiště při práci s infor-



macemi z dané úrovně. První úroveň představuje tok dat (bitstream), který se může skládat na druhé úrovni do souborů (filestream). Závěrečnou úroveň pak tvoří prezentace souhrnného objektu pro úplné zachycení intelektuální entity, typicky se sestávající z množiny souborů.

Důležitou součástí modelu jsou události, které popisují, jaké procesy a činnosti se objektu týkají. Vzhledem k tomu, že na dlouhodobou ochranu může mít vliv celá řada činností včetně změny obsahu objektu, patří sem např. kontrolní součty a ověření integrity, požadavky na šíření a používání nebo přehledy. Události také mají velmi blízko k vazbám, např. požadavek na událost „derivace“ má za následek vytvoření nového objektu a pro dlouhodobou ochranu je důležité tuto vazbu mezi oběma objekty (původní, nový) evidovat. Datový slovník nabízí pro realizaci vazeb několik sémantických jednotek, které zachycují informaci o derivaci, struktuře, závislosti a jiných vazbách. Důležitým aspektem modelu je zásada přijatá pracovní skupinou s označením „1 : 1“. Objekty vytvořené ze stávajících objektů (formou kopie, verze, transformace apod.) se považují za nové objekty spojené s původním objektem událostí a informací o vazbě. Jedním ze závěrů přehledu byl fakt, že datová úložiště často uchovávají objekty ve více kopiích a s ohledem na dlouhodobou ochranu je potřeba mít informaci o každé kopii. Mechanismus vazeb tak umožňuje propojovat objekty bez nutnosti redukce nebo duplikace potřebných metadat o vzniku nového objektu pro dlouhodobou ochranu. Interně mohou systémy uchovávat tyto údaje ve stromové struktuře, v případě výměny dat je ale potřebné předávat pro daný objekt jeho úplná metadata pro dlouhodobou ochranu.

Následuje: testování

V rámci projektu PREMIS byl kladen důraz na opakovaně prověřovanou mezinárodní spolupráci s cílem vyvinout datový slovník určený pro standardizovanou výměnu metadat pro dlouhodobou ochranu digitálních nebo digitalizovaných objektů datových úložišť. Na straně systémů nevyžaduje specifickou architekturu a poskytuje průvodce pro implementaci jádra metadat pro dlouhodobou ochranu. V rámci celkové kooperace byl PREMIS sponzorován OCLC a RLG a oficiální www stránky byly zpřístupněny pod patronací Kongresové knihovny (http://www.loc.gov/premis/). Veškerou projektovou dokumentaci a nové informace lze získat právě na tomto místě. V závěrečné přípravné fázi se nyní nachází poslední cíl projektu: vybudovat pro datový slovník testovací prostředí. Sémantické jednotky datového slovníku byly přepsány do tvaru XML schématu (http://www.loc.gov/standards.premis/schemas.html), jehož otestování proběhne na vybraných nových projektech ve fázi implementace. Také bude prověřena jeho funkce coby výměnného formátu. Lze doufat, že se podobně podaří

zapojit i existující implementace datových úložišť v rámci analýzy a porovnání svých vlastních metadat se sémantickými jednotkami datového slovníku. Předpokládá se, že datový slovník a XML schéma bude stabilní, ačkoliv může docházet k dílčím změnám na základě zkušeností získaných z fáze testování.

B. Další části mozaiky

Jak bylo zmíněno výše, existují další části datového modelu, se kterými metadata pro dlouhodobou ochranu potřebují pracovat a které nejsou součástí datového modelu definovaného pracovní skupinou PREMIS. Například k nim patří metadata práv pro šíření a užívání nebo detailní technická metadata včetně informace o fyzických formátech.

Metadata práv

Metadata práv byla v rámci PREMIS definována zúženě a někdo by mohl namítnout, že i informace o zpřístupňování a šíření je důležitá pro zachování dlouhodobé ochrany. Nicméně nyní probíhá řada aktivit, které se týkají jazyka pro vyjádření práv, a standardizačních pokusů v rámci zpřístupňování a šíření. Mezi hlavní iniciativy na tomto poli lze počítat projekt INDECS Evropské unie, úsilí vydavatelů v rámci ONIX/EDItEUR a Iniciativu pro správu elektronických zdrojů DLF (Digital Library Federation).

Technická metadata

V rámci přehledu, který provedla pracovní skupina PREMIS, se ukázalo velmi časté používání formátu METS coby kontejneru pro seskupení a zabalení metadat digitálního objektu. Typy technických metadat a jejich množství se pak různily a lišily v závislosti na schopnostech datového úložiště automaticky sbírat a ukládat takové údaje. Oblast, kde standardizace technických metadat dosáhla značného pokroku, představují obrazové soubory. Prototyp standardního datového slovníku byl dokončen v rámci NISO v roce 2002 (http://www.niso.org/standards/resources/z39_87_trial_use.pdf). Současně je ale v praxi už široce používán MIX (http://www.loc.gov/mix), rozšiřující schéma formátu METS založené právě na datovém slovníku NISO. Rychlé převzetí tohoto standardu a schématu naznačuje, že datová úložiště se o ukládání technických metadat velmi zajímají a hledají pro ně vhodné vzory a modely. Současně si ale musí být knihovny vědomy toho, že technická metadata je potřeba standardizovat minimálně s ohledem na současný pokrok v průmyslové oblasti vzhledem k tomu, že získávání technických metadat z objektů by mělo probíhat ještě automatizovaněji než v případě extrakce dat pro PREMIS. Na www stránkách formátu METS se nabízí odkazy na několik lokálních schémat pro technická metadata různých typů objektů, které by mohly sloužit jako základ snahám o standardizaci dalších typů ve srovnání s obrazovými objekty (http://www.loc.gov/mets).

Adresáře formátů

Druhým důležitým prvkem metadat pro dlouhodobou ochranu je snadný přístup k údajům o fyzických formátech elektronických souborů. V některých případech lze takovéto údaje najít na www stránkách společností odpovědných za používané formáty (pokud existují), ale je zřejmé, že to není dostatečný způsob jak získat tuto informaci. Z hlediska dlouhodobé ochrany znalost fyzického formátu napomáhá kontrole integrity, umožňuje ocenit riziko spojené s volbou konkrétního formátu a naznačí cestu pro případnou migraci. Díky pochopení fyzického formátu také lze lépe určit možnosti automatické extrakce metadat z digitálního

objektu, a tak jich využít pro naplnění některých údajů z datového slovníku PREMIS nebo vytvoření podrobných technických metadat. Existují dva významné projekty, jejichž cílem je provoz průběžně aktualizovaných registrů formátů budovaných kooperativním způsobem, není ale jasné, zda budou po ukončení dále rozvíjeny. Prvním projektem je PRONOM Národního archivu Velké Británie (http://www.nationalarchives.gov.uk/pronom/), který vznikl jako lokální řešení hledající odpověď na potřebu archivářů bojovat se zastaráváním softwaru formou migrace dokumentů. Na Internetu se poprvé objevil v roce 2004 a v roce 2005 byla nabídnuta podstatně vylepšená verze. S ohledem na veřejné dokumenty je tento adresář zaměřen zejména na textové formáty.

Druhým projektem, který se svou koncepcí prosadil, je adresář GDFR (Global Digital Format Registry, viz. http://hul.harvard.edu/gdfr/), který byl založen z iniciativy DLF v roce 2003. Jakmile byl teoretický model adresáře zpracován na Harvardské univerzitě, jeho prototyp formou služby vznikl na Pensylvánské univerzitě s názvem FRED (Format Registry Demonstration, viz. http://tom.library.upenn.edu/fred/). Jeho prostřednictvím pak mohou správci digitálních objektů experimentovat, do jaké míry lze tuto službu využít a jak ji implementovat. Ačkoliv nejde o příliš atraktivní oblast, její význam je pro dlouhodobou ochranu napříč různými formáty objektů nesporný a vyžaduje kooperativní přístup.

Závěrem

Na základě předchozích koncepčních modelů a získaných zkušeností při jejich implementaci je potřeba pokračovat a krok za krokem se soustředit na nově se objevující vzory a standardy metadat s cílem dlouhodobé ochrany obsahu datových úložišť. Jejich správci a provozovatelé už dnes nemusí stavět na zelené louce. Testovací fáze jádra metadatového formátu PREMIS, pozornost zaměřená na podrobná technická metadata a kooperace v oblasti adresářů fyzických formátů představují témata, kterými bychom se měli zabývat v blízké budoucnosti.

Význam uvedeného příspěvku Sally H. McCallum je nesporný. Dokládá to i rostoucí zájem o dlouhodobou ochranu na pracovních jednání skupiny IT-SDRUK, která byla založena v roce 2005 s cílem podpořit digitalizační procesy v knihovnách. Správci digitálních sbírek a systémů digitálních knihoven (zejména těch v provozu) logicky obracejí pozornost k faktu, že dlouhodobá ochrana se nesestává jen ze zachování vlastního fyzického datového toku, ale ze schopnosti reprezentovat obsah na všech úrovních tak, jak je definuje datový model PREMIS. Podstatným kritériem úspěchu je také cena, kterou je třeba vynakládat na udržování chodu daného datového úložiště až do chvíle, kdy je třeba provést migraci do nové generace. V tom, abychom byli na takovýto proces dobře připraveni a dokázali jej řídit, nám může pomoci právě evidence metadat pro dlouhodobou ochranu. Využití datového slovníku PREMIS pro tento účel se nabízí automaticky. Bylo by tedy velmi žádoucí, aby zejména systémy budované na národní úrovni analyzovaly a vážně zvážily možnost použití formátu PREMIS. Pozorní čtenáři si mohli všimnout v příspěvku Sally H. McCallum opakované zmínky formátu METS. Jeho významu se budu věnovat ve druhé části článku.

METS – výměnný formát pro metadata

Úvod

Formát METS (*Metadata Encoding and Transmission Standard*) je standardem digitálních knihoven druhé generace. Soustředí se na zápis a výměnu všech potenciálních metadat, které se mohou k digitálnímu objektu ve smyslu intelektuální entity vázat, v jednom provedení. Jeho cílem je sloužit jako výměnný formát při přenosu metadat mezi systémy. Jeho reprezentací je XML schéma a správcem formátu je Kongresová knihovna, na jejíž www stránkách lze najít oficiální informace (viz. http://www.loc.gov/standards/mets/). Dne 6. prosince 2005 zde byla uvolněna revize aktuální verze formátu METS 1.5 (http://www.loc.gov/standards/mets/mets-schemadocs.html).

Je potřeba výměnného formátu?

Proč takový formát vznikl? Potřeba formátu METS plyne z obecné vlastnosti správy digitálních objektů a prezentace intelektuálních entit uložených v rámci digitální knihovny. Např. intelektuální entita coby digitalizovaný titul periodika se typicky skládá ze samostatných digitálních objektů (jednotlivých obrázků), které představují strany konkrétního čísla novin nebo časopisu. Každý obrázek musí obsahovat informaci, ke kterému titulu, ročníku, roku, číslu apod. patří a jaká je jeho strana nebo pořadí v souboru všech obrázků. Musí obsahovat informaci, v jakém formátu obrázek je, pro jaký účel má sloužit (zda se jedná o náhled, o archivní nebo uživatelskou podobu), jak je veliký, kdo je oprávněn jej užívat apod. Každý obrázek může být podroben metodě OCR (Optical character recognition), která jej s určitou chybovostí převede do textové podoby a tento indexový soubor může být následně využit pro plnotextové vyhledávání. Každý obrázek může být opatřen analytickým popisem, který poskytne strukturované údaje o jednotlivých publikovaných textech (článcích). Je přitom zřejmé, že jeden článek může být na více stranách stejně jako, že na jedné straně může být více článků. Z uvedeného příkladu je jasné, že souborné objekty mohou existovat na různých úrovních: lze nabízet souborný objekt celého titulu, určitého ročníku, konkrétního čísla nebo článku. Metadata, které digitální knihovna uchovává společně se samotnými objekty, lze řadit do čtyř skupin:

- popisná (abychom mohli objekt snadněji nalézt)
- technická metadata (abychom mohli vyjádřit vlastnosti objektu – např. textový dokument se liší od obrazového apod.)
- strukturální metadata (abychom mohli propojit objekty a metadata mezi sebou i navzájem)
- administrativní metadata (abychom mohli řídit přístup k objektům, zajistit jejich dlouhodobé zachování, zabezpečit dodržení autorského práva apod.)

Protože je budování a provoz digitální knihovny finančně náročný podnik, mělo by být hlavní snahou jejich správců a majitelů používání „stejných“ typů metadat, tj. jejich standardizace. Tento proces se odehrává přirozenou cestou (projekty jdoucí svou vlastní cestou jsou čím dál dražší a nákladnější, než dojde k jejich převodu do standardního prostředí nebo ukončení). Existuje řada již standardizovaných metadatových schémat, zejména v oblasti popisných metadat (MARCXML, MODS, Dublin Core, EAD, TEI). Složitější je situace pro další druhy metadat, kde existuje pouze standard pro technické údaje získávané z obrazových objektů (Z39.87, MIX).

Formát METS se snaží odpovědět na komplexní strukturu intelektuální entity a zachytit její veškerá metadata takovým způsobem, který by umožnil jejich snadnou výměnu mezi různými

systémy podle potřeby. Tam, kde to je možné, využívá existence současných standardů.

Struktura formátu METS

Pro základní popis a bližší seznámení se strukturou formátu METS jsem si dovolil využít a volně přeložit existující základní popis dostupný na www stránkách Kongresové knihovny coby správce formátu (viz. http://www.loc.gov/standards/mets/METSOverview.v2.html):

Dokument formátovaný podle standardu METS se skládá ze 7 částí:

- hlavičky, která obsahuje administrativní informace o METS dokumentu jako takovém (kdo a kdy ho vytvořil, kdo a kdy ho upravil apod.)
- sekce popisných metadat, která mohou být vložena přímo v METS dokumentu a/nebo uvedena odkazem na externí zdroj
- sekce administrativních metadat, která poskytují informace o jednotlivých objektech, o právech pro jejich zpřístupnění a šíření, o původním objektu, který byl vzorem pro digitalizaci apod. (podobně jako u popisných metadat mohou být vložena přímo v METS dokumentu a/nebo uvedena odkazem na externí zdroj)
- sekce souborů, která obsahuje seznam všech fyzických souborů, z nichž se skládá popisovaný objekt nebo objekty, a jejich umístění
- sekce strukturální mapy, která je klíčovou a povinnou částí každého METS záznamu – zachycuje hierarchickou strukturu a vazbu mezi soubory, objekty a metadaty
- sekce strukturálních odkazů, které umožňují odkazovat mezi jednotlivými uzly strukturální mapy, což je velkou výhodou zejména při zachycení struktury složitějších struktur, např. při archivaci www stránek
- sekce pravidel chování, která lze využít pro definici akcí nebo událostí, jež mají nastat při manipulaci s částmi METS dokumentu

Nyní podrobněji rozeberu každou část a uvedu příklad její stručné podoby v XML schématu.

METS hlavička

Hlavička METS dokumentu má za cíl zaznamenat minimální popisná metadata o samotném METS dokumentu. Tyto údaje obsahují datum vytvoření, datum poslední úpravy a status METS dokumentu. Je také možné uvést jména osob nebo institucí, které mají určitý vztah k METS dokumentu, a stručně jej charakterizovat. Nakonec je možné uvést seznam identifikátorů, které slouží pro jednoznačnou identifikaci METS dokumentu.

Ukázka obsahuje dva atributy uvedené přímo v elementu <metsHdr> (CREATEDATE, RECORDSTATUS), který se používá pro záznam data a času vytvoření METS dokumentu a vyjádření, v jakém stavu zpracování se METS dokument nachází. Dále jsou uvedeny 2 subjekty, které mají nějaký vztah k METS dokumentu, jedná se o osoby-jednotlivce a jejich role je podchycena v atributu role.

```
<metsHdr CREATEDATE="2003-07-04T15:00:00" RECORDSTATUS="Complete">
   <agent ROLE="CREATOR" TYPE="INDIVIDUAL">
     <name>Jerome McDonough</name>
   </agent>
   <agent ROLE="ARCHIVIST" TYPE="INDIVIDUAL">
     <name>Ann Butler</name>
   </agent>
</metsHdr>
```

Popisná metadata

Sekce pro popisná metadata může obsahovat jeden nebo více <dmdSec> prvků. Každý <dmdSec> prvek může obsahovat odkaz na externí metadatový záznam (prvek <mdRef>) a/nebo přímo vložená metadata (pomocí prvku <mdWrap>). V případě odkazování na externí metadata se předpokládá použití URI (Uniform Resource Information), které může nabýt těchto typů: URN, URL, trvalé URL, Handle, DOI nebo jiného směrovacího schématu. V případě vkládání lokálních metadat mohou být vnořena pomocí XML schématu nebo pomocí binárního řetězce kódovaného Base64.

U každého záznamu popisných metadat by měl být uveden jejich typ, např.: MARC, MODS, EAD, VRA Core, Dublin Core, NISOIMG (NISO Technical Metadata for Digital Still Images), LC-AV (Library of Congress Audiovisual Metadata), TEIHDR (TEI Header), DDI (Data Documentation Initiative), FGDC (Federal Geographic Data Committee Metadata Standard) nebo jiné metadatové schéma označené jako „OTHER“.

Ukázka popisných metadat volaných externím odkazem a vložených přímo v záznamu:

```
<dmdSec ID="dmd001">
    <mdRef LOCTYPE="URN" MIMETYPE="application/xml" MDTYPE="EAD"
    LABEL="Berol Collection Finding Aid">urn:x-nyu:fales1735</mdRef>
</dmdSec>


<dmdSec ID="dmd002">
   <mdWrap MIMETYPE="text/xml" MDTYPE="DC" LABEL="Dublin Core Metadata">
     <xmlData>
       <dc:title>Alice's Adventures in Wonderland</dc:title>
       <dc:creator>Lewis Carroll</dc:creator>
       <dc:date>between 1872 and 1890</dc:date>
       <dc:publisher>McCloughlin Brothers</dc:publisher>
       <dc:type>text</dc:type>
     </xmlData>
   </mdWrap>
</dmdSec>


<dmdSec ID="dmd003">
   <mdWrap MIMETYPE="application/marc" MDTYPE="MARC" LABEL="OPAC Record">
     <binData>MDI00DdjamOgIDIyMDAlODkgYSAONUOwMDAxMDA...(etc.)
     </binData>
   </mdWrap>
</dmdSec>
```

Každý prvek <dmdSec> musí obsahovat jednoznačný identifikátor (atribut ID). Tím je zajištěno jedinečné označení pro potřeby strukturální mapy, aby bylo možné přiřadit konkrétní metadatový záznam k jednomu nebo více fyzickým souborům.

Administrativní metadata

Prvek <amdSec> může obsahovat administrativní metadata jak fyzických souborů, ze kterých se skládá digitální objekt, tak původních objektů, ze kterých popisovaný objekt vznikl. V rámci METS dokumentu se předpokládají 4 kategorie administrativních metadat:

1. technická metadata (datum a čas vytvoření souboru, fyzický formát)
2. metadata práv a duševního vlastnictví (copyright, licence k užití)
3. zdrojová metadata (popisná a administrativní metadata týkající se vzorového objektu, na jehož základě stávající objekt vznikl)
4. metadata o digitalizaci (vazba mezi zdrojovým a stávajícím objektem, vlastnosti digitalizace)

Každá taková kategorie by měla být zaznamenána v rámci <amdSec> v samostatném dílčím prvku: <techMD>, <rightsMD>, <sourceMD>, and <digiprovMD>. Jejich výskyt je opakovatelný podle potřeby. Ve všech čtyřech kategoriích lze metadatový záznam vložit přímo nebo externím odkazem podle principu prezentovaného v sekci popisných metadat. Důležitou zásadou je zachování jednoznačných identifikátorů pro každý metadatový záznam napříč celým METS dokumentem.

```
<techMD ID="AMD001">
   <mdWrap MIMETYPE="text/xml" MDTYPE="NISOIMG" LABEL="NISO Img. Data">
     <xmlData>
       <niso:MIMEtype>image/tiff</niso:MIMEtype>
       <niso:Compression>LZW</niso:Compression>
       <niso:PhotometricInterpretation>8</niso:PhotometricInterpretation>
       <niso:Orientation>1</niso:Orientation>
       <niso:ScanningAgency>NYU Press</niso:ScanningAgency>
     </xmlData>
   </mdWrap>
</techMD>
```

Příklad prvku <techMD>, který zachycuje technická metadata obrazového souboru:

Pro fyzický soubor, který by měl obsahovat vazbu na tento popisný záznam, by byl v seznamu souborů použit odkaz na hodnotu konkrétní ID příslušného popisného záznamu (AMD001):

```
<file ID="FILE001" ADMID="AMD001">
   <FLocat LOCTYPE="URL">http://dlib.nyu.edu/press/testimg.tif</FLocat>
</file>
```

Sekce souborů

Sekce souborů (prvek <fileSec>) může obsahovat jeden nebo více prvků pro vyjádření skupiny souborů (<fileGrp>). Skupina je sada souborů, které k sobě mají nějaký vztah (např. se může jednat o verze jednoho objektu – obrázek ve formátu JPEG, TIFF a GIF – nebo sekvenční pokračování – audio nahrávka rozdělená do tří částí). Následující příklad představuje jeden objekt (záznam ústní historie) vyjádřený 3 soubory (textový přepis, originální nahrávka ve WAV formátu a nahrávka pro běžný poslech v komprimovaném MP3 formátu):

```
<fileSec>
   <fileGrp ID="VERS1">
     <file ID="FILE001" MIMETYPE="application/xml" SIZE="257537" CREATED="2001-06-10">
       <FLocat LOCTYPE="URL">http://dlib.nyu.edu/tamwag/beame.xml</FLocat>
     </file>
   </fileGrp>
   <fileGrp ID="VERS2">
     <file ID="FILE002" MIMETYPE="audio/wav" SIZE="64232836"
       CREATED="2001-05-17" GROUPID="AUDIO1">
       <FLocat LOCTYPE="URL">http://dlib.nyu.edu/tamwag/beame.wav</FLocat>
     </file>
   </fileGrp>
   <fileGrp ID="VERS3" VERSDATE="2001-05-18">
     <file ID="FILE003" MIMETYPE="audio/mpeg" SIZE="8238866"
       CREATED="2001-05-18" GROUPID="AUDIO1">
       <FLocat LOCTYPE="URL">http://dlib.nyu.edu/tamwag/beame.mp3</FLocat>
     </file>
   </fileGrp>
</fileSec>
```

Výhody použití prvku <fileGrp> se projevují všude tam, kde roste počet jednotlivých souborů, ze kterých se skládá objekt, např. digitalizovaná kniha apod. V případě, že se soubory týkají stejného objektu, lze použít atribut GROUPID. Každý fyzický soubor musí mít jedinečný identifikátor, který zajišťuje korektní použití vazeb v strukturální mapě METS dokumentu.

Vlastní obsah fyzického souboru může být odkazován formou externího umístění (prvek <FLocat>) nebo přímým vložením do METS dokumentu (pomocí prvku <FContent>). Druhý způsob pak umožňuje využít METS jako výměnný formát nejen pro metadata, ale i pro přenos vlastních souborů.

Strukturální mapa

Strukturální mapa METS dokumentu (prvek <structMap>) obsahuje hierarchicky definovanou strukturu pro prezentaci a zachycení vztahů uvnitř METS dokumentu mezi jeho jednotlivými částmi (metadata a fyzické soubory). Každou vazbu lze vyjádřit pomocí opakovaně vnořených prvků . V jeho rámci pak lze použít prvky <mptr> a <fptr> pro odkazy na relevantní část METS dokumentu, resp. konkrétního fyzického souboru. Následuje příklad jednoduché strukturální mapy:

```
<structMap TYPE="logical">
   <div ID="div1" LABEL="Oral History: Mayor Abraham Beame"
     TYPE="oral history">
     <div ID="div1.1" LABEL="Interviewer Introduction"
       ORDER="1">
   <fptr FILEID="FILE001">
     <area FILEID="FILE001" BEGIN="INTVWBG" END="INTVWND"
       BETYPE="IDREF" />
   </fptr>
   <fptr FILEID="FILE002">
     <area FILEID="FILE002" BEGIN="00:00:00" END="00:01:47"
       BETYPE="TIME" />
   </fptr>
   <fptr FILEID="FILE003">
     <area FILEID="FILE003" BEGIN="00:00:00" END="00:01:47"
       BETYPE="TIME" />
   </fptr>
</div>
```

```
    <div ID="div1.2" LABEL="Family History" ORDER="2">
    <fptr FILEID="FILE001">
      <area FILEID="FILE001" BEGIN="FHBG" END="FHND"
        BETYPE="IDREF" />
    </fptr>
    <fptr FILEID="FILE002">
      <area FILEID="FILE002" BEGIN="00:01:48"END="00:06:17"
        BETYPE="TIME" />
    </fptr>
    <fptr FILEID="FILE003">
      <area FILEID="FILE003" BEGIN="00:01:48" END="00:06:17"
        BETYPE="TIME" />
    </fptr>
  </div>
    <div ID="div1.3" LABEL="Introduction to Teachers' Union"
      ORDER="3">
    <fptr FILEID="FILE001">
      <area FILEID="FILE001" BEGIN="TUBG" END="TUND"
        BETYPE="IDREF" />
    </fptr>
    <fptr FILEID="FILE002">
      <area FILEID="FILE002" BEGIN="00:06:18" END="00:10:03"
        BETYPE="TIME" />
    </fptr>
    <fptr FILEID="FILE003">
      <area FILEID="FILE003" BEGIN="00:06:18" END="00:10:03"
        BETYPE="TIME" />
    </fptr>
  </div>
```

Strukturální mapa popisuje záznam ústní historie (se starostou New Yorku Abraham Beame) obsahující 3 části: úvod, rodinné pozadí starostovy rodiny a rozhovor na téma učitelských odborů. Každá část pak obsahuje vazby na 3 fyzické soubory (text s přepisem, primárná audio nahrávku a audio nahrávku v uživatelském formátu), jejichž záznam jsme si představili výše. Prvek <fptr> zde zastupuje vazbu s pomocí identifikátoru každého fyzického souboru.

Strukturální odkazy

Sekce strukturálních odkazů umožňuje pomocí prvku <smLink> zaznamenat existující odkaz mezi jednotlivými částmi METS dokumentu (např. mezi jednotlivými částmi strukturální mapy). Použití strukturálních odkazů lze demonstrovat na příkladu METS dokumentu pro www stránku s obrázkem, který vede jako odkaz na jinou www stránku. Nejprve je uvedena strukturální mapa, poté strukturální odkaz:

```
<div ID="P1" TYPE="page" LABEL="Page 1">
  <fptr FILEID="HTMLF1"/>
     <div ID="IMG1" TYPE="image" LABEL="Image Hyperlink to
       Page 2">
     <fptr FILEID="JPGF1"/>
</div>

<div ID="P2" TYPE="page" LABEL="Page 2">
  <fptr FILEID="HTMLF2"/>
</div>
```

```
<smLink from="IMG1" to="P2" xlink:title="Hyperlink from
JPEG Image on Page 1 to Page 2" xlink:show="new"
xlink:actuate="onRequest" />
```

Sekce pravidel chování

V případě, že od objektu požadujeme určité chování, lze je definovat v sekci pro pravidla chování. Ta může obsahovat jeden nebo více prvků <behavior>, z nichž každý obsahuje jednak prvek pro definici rozhraní, které představuje souhrn pravidel chování, jednak prvek pro mechanismus, kterým se konkrétní pravidlo chování realizuje. Následující příklad uvádí odkaz na distribuovanou www službu poskytovanou v rámci projektu Fedora:

Na výše uvedené struktuře formátu METS bylo cílem ukázat, že poskytuje jednoduchý a přitom škálovatelný způsob zápisu všech druhů metadat, které se týkají prezentace a uložení intelektuální entity reprezentované jedním nebo více digitálními objekty. Díky tomu jej lze efektivně využít jako výměnný formát. Pro jeho praktické využití je ale potřeba zmínit ještě některé další informace – existenci METS profilů a METS rozšíření, které usnadňují praktické nasazení METS formátu.

METS profily a adresář implementací

Protože formát METS typicky nedefinuje (s výjimkou rozšíření, viz. dále) jednotlivá schémata pro popis jednotlivých druhů metadat, předpokládá, že každá implementace o sobě bude publikovat tzv. METS profil. Ten slouží k informaci správců datových úložišť o tom, co mohou od sebe vzájemně očekávat. Způsob, jak lze profil METS vytvořit, je uveden na www stránce http://www.loc.gov/standards/mets/mets-profiles.html společně s příslušným XML schématem. V současné době jsou zveřejněny profily *Oxford Digital Library*, *University of California – Berkeley*, projektu Greenstone, Kongresové knihovny a profil RLG. Jde o profily digitálních knihoven, které jsou v provozu. Úplný přehled implementací METS formátu lze najít na samostatné www stránce (viz. http://sunsite.berkeley.edu/mets/registry/), kde jsou uvedeny kromě fáze, ve které se projekt nachází, také kontaktní údaje, záběr, druh zpracovávaných objektů, předpokládaný výstup a informace, zda byl vyvinut speciální software, který by mohl být prospěšný ostatním. Mezi seznamem institucí jsou např. italská *Biblioteca Digitale Provinciale P. Albino*, portugalská *Biblioteca Nacional*, čínské ministerstvo vzdělávání, *Culturnet Cymru* a Národní knihovna Walesu, německé *Göttinger Digitalisierungs-Zentrum*, anglická *Oxford University*, rakouská *University of Graz* aj. Postupně se tak světové zastoupení vyrovnává původní převaze implementací ze Spojených států.

Rozšiřující schémata METS

Vzhledem k potřebě zápisu administrativních metadat, která dosud nebyla žádným způsobem standardizována, existuje v rámci standardu METS možnost definovat vlastní rozšíření. Na základě praktických zkušeností tak byla už na samotném začátku existence formátu METS definována některá schémata, kterých mohou správci digitálních knihoven využít spíše než se pouštět do definicí vlastních metadatových formátů. Jedná se o:

- schéma pro technická metadata textových objektů (*New York University*)
- schéma pro technická metadata audiovizuálních objektů (*Library of Congress*)
- schéma pro technická metadata obrazových objektů (MIX, na základě NISO Z39.87)
- schéma pro administrativní metadata práv

Závěrem

Systémy digitálních knihoven se posouvají do své druhé generace. Větší důraz je kladen na užívání objektů bez manuálního zásahů správců digitálních knihoven a datových úložišť a lze přepokládat rozmach přebírání a sdílení digitálních objektů. Předpokladem je standardizace metadatových schémat a usnadnění jejich sdílení a přebírání. Výměnný formát METS tuto funkci plní velmi dobře a poměrně rychle se prosazuje ve stávajících i nových implementacích. Role nových standardů bezpochyby usnadňuje nákladnou cestu digitálních knihoven a umožňuje rychlejší adaptaci na změny v prostředí. Analýzu jejich implementace lze tedy pro české projekty doporučit co nejdříve.