Ostráková, Natalie; Kočišová, Pavlína; Beňačková, Miroslava

Vývoj standardu PREMIS a možnosti jeho dalšího využití ve standardech NDK

ProInflow. 2019, vol. 11, iss. 2, pp. [72]-85

ISSN 1804-2406 (online)

Stable URL (DOI): https://doi.org/10.5817/ProIn2019-2-6
Stable URL (handle): https://hdl.handle.net/11222.digilib/141802

Případová studie

ProInflow: časopis pro informační vědy
Vol. 11, No. 2, 2019, str. 72–85
DOI: https://doi.org/10.5817/ProIn2019-2-6

http://www.phil.muni.cz/journals/index.php/proinflow

VÝVOJ STANDARDU PREMIS A MOŽNOSTI JEHO DALŠÍHO VYUŽITÍ VE STANDARDECH NDK

DEVELOPMENT OD THE PREMIS STANDARD AND THE POSSIBILITIES OF ITS FURTHER APPLICATION IN NDL STANDARDS

Natálie Ostráková, Pavlína Kočišová, Miroslava Beňačková

Národní knihovna ČR

Abstrakt

Účel – Článek představuje metadatový standard PREMIS, který se používá pro zápis archivačních metadat, je diskutována historie standardu, struktura a možnosti jeho užití v různých institucích a aplikacích. Primárně ale přibližuje jeho implementaci v standardech Národní digitální knihovny České republiky (NDK). Cílem je poukázat na jeho výhody a na směry, jakými by se mohl Standard NDK v oblasti zápisu archivačních metadat ubírat. Zařazením se jedná o případovou studii.

Design/metodologie/přístup – Metodologicky se jedná o analýzu. V její úvodní části je shrnuta historie standardu PREMIS, jeho části a proměny. Následně se článek věnuje konkrétním detailům aplikace standardu PREMIS v NDK, používaným elementům a řízeným slovníkům. Taktéž se zmiňuje způsob aplikace PREMIS v zahraničních institucích. V poslední části se potom nastiňuje výhled užití PREMIS v NDK a možné implementace jeho dalších funkcí, což je ukazatelem, jakým směrem se bude vývoj v NDK ubírat.

Výsledky – Analýza ukázala způsoby užití standardu v některých zahraničních institucích, některé z nich jsou zvažovány v budoucím rozvoji Standardu NDK. Jedná se například o možnost popisu na úrovni reprezentace a popisu intelektuální entity jako objektu dlouhodobého uložení; nebo o přidání elementu rozšíření u objektu pro zachycení co nejkomplexnějších technických vlastností.

Originalita/hodnota – Přínosem článku je ucelený pohled na využití mezinárodního standardu PREMIS ve Standardu NDK a naznačení dalšího možného vývoje. Spolu s tím jsou diskutována rozhodnutí, která za těmito změnami stojí.

Klíčová slova: PREMIS, Národní digitální knihovna, archivační metadata, digitální archivace

Abstract

Purpose – This article presents the metadata standard PREMIS, which is used for storage of archival metadata, history, structure and uses in other institutions are discussed. Main focus is on implementation of the standard within the standards of the National Digital Library of the Czech Republic (NDL). Its goal is to point out PREMIS´s advantages and show directions, in which the NDL standards could be upgraded in the field of preservation metadata. It can be classified as a case study.

Design/Methodology/Approach – Regarding used methodology, this article is an analysis. The history of PREMIS standard, its parts and modifications are described in the opening part. Next, it discusses specific details, used elements and controlled vocabularies applied to the NDL standard. It also mentions different PREMIS applications in the foreign institutions. The last section deals with the future use of PREMIS within NDL and possible implementations of its other functions – which is a good way to indicate further direction of NDL.

Results – The analysis has shown implementation of the standard in some institutions, some of them are considered in the next development of the national standard. For example, the possibility of the description on the representation level and the description of the intellectual entity as a long-term preservation object; or adding the extension element to Object description for storing the most complete technical metadata possible.

Originality/Value – The benefit of the article is a comprehensive overview of the use of international standard PREMIS in national standard with a hint of future development. Together with it some decisions about the changes are discussed.

Keywords: PREMIS, National Digital Library, preservation metadata, digital archiving

Úvod

Pro efektivní správu a využívání nejen digitálních dokumentů jsou klíčovým prvkem metadata, jež spravovaná data popisují. V běžné knihovnické praxi jsou nejužívanějšími popisná metadata, založená na katalogizačních standardech, primárně určených pro popis tištěných knih a časopisů.

Digitalizace a vznik digitálních dokumentů s sebou přináší problematiku dlouhodobého uchovávání digitálních objektů, kde přibyla potřeba zaznamenávání dalších typů metadat. Přístupnost k digitálním objektům je totiž komplikovanější; nelze k nim přistupovat přímo a musí být tedy doplněny dostatečnými informacemi pro současné i budoucí využívání zde uložených informací.

Archivační metadata, jimž je tento článek věnován, mají podpořit dlouhodobou přístupnost a použitelnost digitálních dokumentů zaznamenáním informací o klíčových vlastnostech archivovaného objektu, o jeho životním cyklu, (tj. použitém hardware a software), případně i o informacích, nutných k otevření souboru. Na rozdíl od popisných metadat, která jsou potřebná a využitelná v době svého vzniku, (tj. plní funkce a potřeby, které máme nyní), vznikají archivační metadata zřejmě i dlouho předtím, než budou využívána. Je tedy nutné při jejich tvorbě přemýšlet o budoucích potřebách při archivaci digitálních dokumentů. (Dappert, 2018). Právě proto implementace archivačních metadat nebývá jednoduchá, a ne vždy jsou tyto informace zaznamenány standardizovaně, či vůbec zaznamenávány. Pro standardizovaný zápis archivačních metadat byl pak vytvořen standard PREMIS, který je aktuálně implementován v komerčních i volně dostupných archivačních řešeních (Archivematica, Roda, Rossetta) a je využíván v institucích po celém světě (Kongresová knihovna, Národní digitální knihovna Finska).

Standard PREMIS

Standard PREMIS byl vytvořen v roce 2005 a je založen na výstupu mezinárodní pracovní skupiny The OCLC/RLG Working Group on Preservation Metadata, která fungovala mezi lety 2002-2005. Tato skupina rozpracovala kategorie informačních objektů modelu OAIS – interpretační informace a archivační informace do systému metadatových elementů. OAIS Premisu poskytl koncept taxonomie informačních objektů, balíčků pro archivované objekty i strukturu připojených metadat. Z modelu OAIS Premis také přebírá koncept vztahů mezi různými druhy informací, které jako celek identifikují dokument (provenienční a kontextuální informace). Tento základ byl v následujících letech redigován a rozšiřován až do aktuální verze 3.0, vydané v roce 2015.

Standard PREMIS je spravován pomocí orgánů Managing Agency a Editorial Comitee, které spolupracují s PREMIS Implementation Group (PIG) v rámci organizované údržby a aktualizace *(Maintenance activity)*. Standard je podporován Kongresovou knihovnou, která jeho údržbu také hradí. Hlavními výstupy jsou Datový slovník *(Data Dictionary)* a xml schéma PREMIS. Datový slovník obsahuje komplexní sadu metadatových elementů, nazývaných jako sémantické jednotky. Tyto jsou samostatně popisovány v krátkých kapitolách, ve kterých je uvedena definice, návod k jejich vyplňování a příklady. Vybrané hodnoty jsou pak obsaženy v řízeném slovníku, spravovaném Kongresovou knihovnou.

Klíčovým prvkem metadatových standardů je sada elementů, která slouží k popisu informačních zdrojů. Definuje se význam elementů, vztahy mezi elementy a dává se návod k jejich plnění. V případě standardu PREMIS nejsou definovány přímo metadatové elementy, ale sémantické jednotky *(Semantic Units)*, přičemž sémantická jednotka se zde chápe jako část informace nebo vědomosti, která má být o objektu zachycena. Naproti tomu, metadatový element je definovaný způsob reprezentování informace v metadatovém záznamu, schématu či databázi. (Caplan, 2017)

V první verzi standardu PREMIS 1.0 byl vytvořen model pro pět základních entit: intelektuální entitu (Intellectual Entity), objekt (Object), činitel (Agent), událost (Event) a práva (Rights). Následující verze 2.0 z roku 2008 již reagovala na potřeby odborné komunity, které vyústily ve čtyři větší změny:

1. Struktura vztahů mezi entitami v datovém modelu byla zobecněna tak, aby vždy zahrnovala obousměrné vazby ve všech případech. To znamená, že vazby mohou být dokumentovány mezi jakýmikoliv dvěma entitami, se kterými PREMIS pracuje.
2. Práva byla rozšířena tak, aby umožňovala větší variabilitu v popisu právních prohlášení, včetně schopnosti zaznamenat informace, specifické pro intelektuální vlastnictví ustanovené copyrightem, licencí či statutem.
3. Datový slovník PREMIS byl doplněn o možnost strukturovaného popisu signifikantních vlastností a úrovní ochrany.

4. Bylo představeno PREMIS xml schéma jako nástroj podpory externích formátů metadat. Zápis pomocí xml schématu je nyní nejužívanějším způsobem zápisu archivačních metadat dle standardu PREMIS. (Dappert, 2018; Lavoie & Gartner, 2013)

Verze PREMIS 2.0 byla třikrát v menší míře doplněna; v případě aktualizace 2.1 (leden 2011) o rozšíření sémantických jednotek pro činitele. Možnosti rozšíření formátu byly nasměrovány blíže k možnostem rozšíření schématu METS. V další verzi 2.2 (červenec 2012) byla více posílena práva *(Rights)* díky přidání nových sémantických jednotek. Úpravou také prošlo schéma PREMIS xml. (Lavoie & Gartner, 2013) Verze 2.3 ze srpna 2014 upravovala pouze xml schéma a upřesnila některé jeho elementy.

PREMIS ve verzi 3.0 z roku 2015 formuluje změny v datovém modelu, z původních pěti entit jej redukuje na čtyři – Objekt, Činitel, Událost a Právní deklaraci. V této nejnovější verzi PREMIS se intelektuální entita (IE) přesunula do podkategorie objektu a netvoří již samostatnou kategorii.

Tato změna byla provedena na základě praxe, konkrétně vytváření agregací (sbírky či vyjádření podle FRBR) zachycovaných metadat a z důvodu verzování událostí aktualizace metadat na úrovni IE. (Cubr, 2017-2018)

V této verzi standardu přibyla možnost popsat prostředí *(Environments)*, tedy hardware a software potřebný k užití digitálních objektů; dále je možné popsat fyzické objekty jakožto objekt. Objekt jako takový dostal novou sémantickou jednotu „preservationLevelType“ pro označení typu uchovávání, které bude užito. Novou sémantickou jednotku dostali také činitel a událost. Činitel má nyní pro záznam verze softwaru element „agentVersion“. Událost získala sémantickou jednotku eventDetailInformation pro připojení externího metadatového schématu pro detailnější popis událostí. (Cubr, 2017-2018)

Datový model PREMIS 3.0 definuje jednotlivé entity takto:

1. Objekt je diskrétní jednotka informace, která je předmětem digitální archivace. Jako objekt může být definováno i prostředí, technologie, která slouží ke spuštění nebo reprodukci objektu.
2. Událost je činností, zahrnující nebo zasahující jeden objekt či činitele, který je spojený s repozitářem.
3. Činitel je osobou, softwarovým programem/systémem nebo organizací, spojený s událostí v životním cyklu objektu, popřípadě s právy, spojenými s objektem. Jako činitel může vystupovat i prostředí.
4. Právní deklarace je prohlášení o jednom nebo více právech, které se vztahují k objektu nebo činiteli.

Prostředí v tomto datovém modelu není samostatnou entitou, jelikož spadá pod objekt, nicméně i jako jeho podřízená jednotka může vytvářet specifické vztahy (může být agentem, co například objekt být nemůže.

Aby bylo možné standard PREMIS využít také k propojování dat z různých zdrojů (tzv. Linked Data), byla vytvořena PREMIS OWL Ontologie (*PREMIS OWL Ontology)*. Jedná se o přepis datového modelu pro datový slovník PREMIS do kódování RDF, přičemž jsou v něm stanoveny třídy a vlastnosti, pomocí kterých lze popsat archivační metadata a vztahy mezi jednotlivými součástmi. Kromě vlastních využívá i pojmy z již existujících ontologií (jako například Dublin Core metadata terms, Provenance Interchange Ontology – PROV-O nebo Simple Knowledge Organization System – SKOS) a podobně jako v datovém slovníku je možné využít hodnoty z kontrolovaných slovníků.

Původně byla tato ontologie vytvořena podle Datového slovníku PREMIS verze 2.2, v současné době již prošla aktualizací podle verze 3.0.

Signifikantní vlastnosti

Kromě výše uvedených konceptů PREMIS řeší také autenticitu objektů. Autenticita je jeden z klíčových pojmů digitální archivace a udržení autenticity je nejvyšším cílem digitální archivace (Caplan in Cubr, 2017, s. 5). Autentický dokument je "právě tak spolehlivý, jako byl spolehlivý v době svého vzniku" (Cubr, 2017, s. 76), tj. nebyl mezitím nijak neplánovaně změněn a každá případná změna (např. migrace formátu) je deklarována. Pro zachování a prokazování autenticity dokumentu jsou důležité také signifikantní vlastnosti, resp. jejich zachování. Koncept signifikantních vlastností (*significant properties*, překládáno též jako „významné vlastnosti") byl představen v aktualizované verzi PREMIS 2.0. Za signifikantní vlastnosti jsou v PREMIS považovány takové charakteristiky objektu, které by měly být uchovány napříč archivačními opatřeními, aby objektu zůstala zachována autenticita.

Lze tedy říci, že signifikantní vlastnosti jsou pokusem o řešení problému ztráty některých vlastností v průběhu archivačního procesu informačního obsahu. Odvozování – transformace digitálního objektu – způsobuje na bitové úrovni změny digitálního objektu, což je jádrem archivačního procesu, který počítá s tím, že v průběhu času bude potřeba transformovat původní digitální datový objekt a jeho obsah (v balíčku SIP, resp. AIP) první verze. Každá transformace potom vyvolává změnu v původních bitech, čímž je vyvolána pochybnost o autenticitě (neporušení) původního informačního obsahu. (Cubr, 2017-2018)

Signifikantní vlastnosti tedy musí být definovány hlavně na úrovni intelektuální entity jakožto vlastnosti toho, co může čtenář vidět v uživatelské aplikaci digitální knihovny. Definice klíčových vlastností intelektuální entity stanoví cíle, kterých je nutno dosáhnout v digitalizaci i při dlouhodobém uchovávání a zpřístupňování, přičemž jejich popis musí být srozumitelný celé cílové čtenářské komunitě jako deklarace toho, co je možné od digitalizovaných dokumentů očekávat. (Cubr, 2017-2018)

Metadatové schéma PREMIS a Standard NDK

Jak bylo výše uvedeno, jsou archivační metadata považována za důležitou součást v procesu dlouhodobého uchovávání dokumentů, podílejí se na zajištění přístupnosti dat v budoucnu. Také Standard Národní digitální knihovny (dále NDK) používá toto schéma pro zápis archivačních metadat.

Konkrétně jej používá pro popis technických vlastností archivovaných objektů a pro popis původu (tj. digital provenance) těchto digitálních objektů. Standard NDK tedy pracuje s následujícími částmi standardu PREMIS: PREMIS Object, PREMIS Event a PREMIS Agent.

PREMIS Object

Objekty jsou tím, co jednotlivé repozitáře spravují a archivují, jsou předmětem dlouhodobé ochrany. Schéma PREMIS ve verzi 2.2 rozlišuje několik úrovní objektů, jenž je možné popsat. Jsou to objekty soubor *(File)*, bitstream, reprezentace *(Representation)* a intelektuální entita *(Intellectual Entity)*. Standard NDK aktuálně počítá s popisem úrovně file, tj. jednotlivých počítačových souborů (např. obraz JPEG2000, soubor ALTO). V případě Standardu pro popis elektronických publikací je možné dobrovolně popsat i úroveň bitstream, tj. objekt, který je součástí objektu file (souboru) a jedná se o data, která nemohou existovat samostatně mimo soubor (např. JPEG2000 vložený v souboru PDF). Pro účely dlouhodobé archivace se zdá užitečné či dokonce potřebné tyto části souborů popsat, ale popis na úrovni bitstream se může ukázat jako časově a výpočetně náročný a zřejmě ne vždy existujícími nástroji možný, proto je zatím tento popis dobrovolný a bude testován v pilotním provozu.

Standard NDK nevyžaduje popis všech objektů, jež jsou součástí informačního balíčku, který se ukládá v úložišti Národní knihovny. Popisují se pouze objekty, jež jsou určeny k dlouhodobému uchovávání (a u kterých existuje možnost, že nad nimi v budoucnu budou prováděny ochranné aktivity) a jejich přímí předchůdci, kde je snahou zaznamenat kompletní životní cyklus objektu. Popisují se tedy archivační obrazy, obrazy, jež jim předcházely (např. primární sken ve formátu TIFF) a soubory ALTO. U jednotlivých objektů se zaznamenávají tyto jejich vlastnosti: jméno a verze souborového formátu, velikost souboru, aplikace, jíž byl soubor vytvořen, kontrolní součty, vztahy k jiným popisovaným objektům a událostem. Standard NDK pro popis objektů používá všechny elementy (se sémantickými jednotkami se ve Standardu NDK pracuje jako s elementy metadatového schématu), které schéma PREMIS určil jako povinné. V případě elektronických publikací využívá Standard i možnosti vložit do schématu PREMIS jiné externí metadatové schéma, které dokáže popsat specifické vlastnosti popisovaného objektu.

PREMIS Event

V části PREMIS Event se ve Standardu NDK zachycují události, které před přijetím do repozitáře objekt nějak mění, tj. vznik objektu (digitalizace, vygenerování XML), migrace do jiného formátu, smazání apod. Popis událostí, jejichž záměrem není objekt jakkoliv měnit (identifikace, validace apod.), se ve Standardu NDK aktuálně nevyžaduje. Standard NDK pro popis událostí používá téměř všechny elementy, které schéma PREMIS pro události předepisuje. Každá událost je zde jednoznačně identifikována pro následné odkazování událost-činitel-objekt. Dále se uvádí typ události, čas, kdy k události došlo, výsledek události a odkazuje se na činitele události a zasažený objekt.

PREMIS Agent

Činitel události je popsán v části PREMIS Agent a je zde možné popsat software, hardware, osobu i organizaci. Uvádí se jméno činitele (název softwaru, příjmení osoby) a v případě migračního softwaru vyžaduje Standard NDK i zápis příkazu k migraci. V případě standardu pro zvukové dokumenty se do části PREMIS Agent vkládá navíc externí metadatové schéma, jenž umožňuje popsat činitele a výrobu objektu podrobněji (výrobce, sériové číslo, nastavení zařízení při digitalizaci).

Výše uvedené má směřovat k tomu, že z metadatového záznamu má být vždy možné zjistit, co se s objekty dělo, jaké akce byly provedeny a jaké nástroje byly použity. To umožňuje kdykoliv později identifikovat "zasažené" objekty, například zjistí-li se chyba ve vytvářejícím softwaru, je možné na základě zaznamenaných informací vyhledat další zasažené objekty v repozitáři. Z tohoto důvodu je dobrou praxí zaznamenání všech dílčích aplikací, jež se na změně objektu podílejí, a to i v případě, kdy jsou jednotlivé aplikace součástí souhrnné aplikace. Zároveň je možné na základě těchto informací možné později prokázat autenticitu objektu.

Kontrolované slovníky

V některých případech je vhodné při vyplňování hodnot elementů využívat kontrolované slovníky. Výhodou takového postupu je následně možnost zautomatizování některých kroků v repozitáři. V datovém slovníku schématu PREMIS je tak všude, kde to tvůrci považovali za vhodné, doporučeno využívat kontrolovaných slovníků. Předpokládá se, že správci repozitářů si takové slovníky vytvářejí a spravují sami. V některých případech však Datový slovník PREMIS navrhuje vhodné hodnoty a pro některé elementy disponuje i doporučeným kontrolovaným slovníkem. Kontrolované slovníky existují pro zapsání typu události, typu agenta, výsledku události, vztahů mezi objekty a jsou dostupné na webové službě spravované Kongresovou knihovnou (http://id.loc.gov/). Uvedené slovníky obsahují hodnoty, které jsou dle autorů aplikovatelné na jakýkoliv repozitář a na žádost je možné je oficiálně o hodnoty doplnit. Repozitáře si pak lokálně mohou tyto kontrolované slovníky přizpůsobit svým potřebám - např. nějaké hodnoty přidat a jiné odebrat. V případě, že repozitář kontrolované slovníky používá, se doporučuje i v metadatovém záznamu uvést, o jaký slovník se konkrétně jedná; konkrétně kvůli případnému exportu dat a jeho následnému znovuvyužití v jiných systémech. (PREMIS Editorial Committee, 2012 s. 18-22).

Zvláště významný je například kontrolovaný slovník typů událostí (Preservation Events Controlled Vocabulary). Nová verze tohoto slovníku (dostupné zde: https://www.loc.gov/standards/premis/v3/preservation-events-revision1.pdf) vyšla v srpnu 2017 a aktuálně obsahuje 42 událostí. Jedná se o události, které mohou mít dopad na dlouhodobou archivaci objektů, tj. především události, které objekty nějak mění, ale i události důležité z hlediska správy objektů v repozitáři (identifikace, validace apod.). Je možné říci, že tento slovník má i edukační funkci, jelikož

uvádí události, které jsou z pohledu odborníků pro dlouhodobé uchovávání důležité a je tak možné jej považovat i za jakýsi manuál akcí, jež v průběhu uchovávání bude třeba provádět.

Standard NDK v případě plnění elementu typ události vyžaduje kontrolovaný slovník, ale konkrétní slovník neuvádí. Jsou zde sice hodnoty, které musí být povinně zaznamenány (capture, migration, derivation, deletion), nicméně formulace uvedená ve Standardu spíše indikuje, že je možné vedle těchto hodnot použít i jiné. Bohužel data uložená v úložišti Národní knihovny ne vždy povinné události popisují. Velmi často chybí událost „deletion", tj. smazání primárního skenu. Tato událost se může zdát nadbytečná, protože smazání primárního skenu nevede ke změně ostatních objektů. Jedná se však o důležitý údaj pro následné využívání dat, například export či agregace, díky které se uživatel dozví, že tento objekt byl smazán a v balíčku tedy není. Naopak se v datech vyskytují i události navíc, například ořez obrazu. Kontrolovaný slovník událostí PREMIS událost ořez neobsahuje, ale obsahuje událost modifikace, pod kterou lze ořez a další úpravy zahrnout. Správný postup v případě zaznamenání události ořezu by tak spíše byl použít „modification" jako typ události a do dalšího elementu, jež obsahuje detailnější informaci o události (eventDetail) uvést, že se jednalo o ořez.

Standard NDK doporučuje využít kontrolované slovníky při plnění více elementů, ovšem ne vždy dává kontrolovaný slovník k dispozici či není ze standardu zřejmé, že hodnoty, které uvádí, jsou jediné možné. Tvorba kontrolovaných slovníků pro Standard NDK je tak dalším úkolem při rozvoji standardizace v rámci NDK.

Uložení metadat PREMIS v SIP balíčku

Existuje několik způsobů, jak PREMIS metadata uložit. Jedním z nich je uložení PREMIS metadat do kontejnerového metadatového schématu METS, což využívá i Standard NDK. Tento postup je schválený Kongresovou knihovnou a schéma PREMIS je i registrovaným schváleným externím metadatovým schématem pro použití v metadatovém schématu METS. Jsou-li metadata PREMIS vložena do schématu METS, pak se umisťují do části amdSec, která je určena pro zaznamenání administrativních metadat. V části amdSec je pak doporučeno vkládat technická metadata týkající se objektů (např. TIFF, JPEG2000) do části techMD a metadata týkající se událostí a činitelů, tj. metadata dokumentující životní cyklus objektů se vkládají do části digiprov (Guidelines for using PREMIS with METS for Exchange, 2017). Tento postup dodržuje i Standard NDK.

Metadata PREMIS je možné uložit i mimo schéma METS, a to například do samostatného metadatového souboru. Tento postup je zvolen například pro projekt E-ARK, kde jsou metadata PREMIS uložena mimo ostatní metadata a jsou z hlavního METS záznamu odkazována (Bredenberg, 2019, s. 28).

Schéma PREMIS ve standardech jiných institucí

Švédská národní knihovna na svém webu zveřejnila doporučení pro tvorbu informačních balíčků (SIP). Dokumentace je z roku 2013 a využívá metadatová schémata, jež jsou využívána i v rámci projektu NDK, jedná se o: METS, MODS, PREMIS, MIX, ALTO. V této dokumentaci se předepisuje použití pouze části

standardu PREMIS, jenž se věnuje popisu objektů (National Library of Sweden, 2013). Oproti Standardu NDK, jenž popisuje pouze úrovně file a bitstream, se zde popisuje i úroveň reprezentace. Tuto úroveň popisu používají pro zaznamenání metadat, jež se týkají více objektů v balíčku (obvykle více objektů stejného typu) a nemusí tedy taková metadata uchovávat u každého jednotlivého objektu, čímž zmenší objem metadat.

Podobný přístup zřejmě zvažuje i softwarová aplikace Archivematica. Standard PREMIS je zde součástí metadatového schématu METS (METS AIP) a slouží k popisu archivovaných objektů, zaznamenání událostí, jež se dějí s objekty a k zaznamenání souvisejících činitelů. Tradičně se zřejmě využívala též úroveň popisu „file“, jenž měla za následek v případě velkých balíčků SIP s tisíci soubory i ohromné množství metadat, které bylo následně problematické indexovat, ukládat a parsovat. Cílem tedy bylo metadatové záznamy zmenšit tím, že se omezí opakování stejných informací, aniž by však došlo ke ztrátě těchto informací. Soubory stejného typu jsou totiž obvykle podrobeny stejným událostem se stejnými agenty a se stejnými výsledky. Možným řešením je použít úroveň popisu representation, s níž by bylo možné popsat všechny události a související agenty týkající se jedné reprezentace intelektuální entity, a tedy více souborů společných vlastností (např. archivačních kopií), tj. vytvoří se událost k jedné reprezentaci, a nikoliv ke každému souboru reprezentace. Události jako ingestion (příjem), message digest calculation, virus scan, format identification by tak metadatový záznam obsahoval pouze jedenkrát pro stejný typ objektů. Jak autoři uvádí, je však třeba následně správně ošetřit různé výsledky událostí, například v případě události validace – některé objekty budou validní, některé nevalidní. (PREMIS/METS for scalablity, 2018). Oproti standardizaci NDK je v systému Archivematica využíván element pro rozšíření u objektu (objectCharacteristicsExtension), kam se ukládá výstup z nástroje FITS, jenž slouží k charakterizaci souborů. (PREMIS metadata: original files, 2011). Zajímavým řešením, které Archivematica dle své dokumenace používá, je popis některých činitelů (název aplikace, verze, příkazový řádek) přímo v části pro událost, nikoliv v části pro činitele. Z dokumentace k systému Archivematica je možné vydedukovat, že popis činitele u události je aplikován na dílčí software, který akci přímo provedl (program pro kontrolu kontrolních součtů souborů, validátor) a zvlášť v části PREMIS Agent je popsán systém Archivematica, který tyto dílčí aplikace obsahuje. Od verze 1.7 systému Archivematica je implementován PREMIS 3.0.

Finská národní digitální knihovna také pracuje se schématem PREMIS. Oproti Standardu NDK doporučují využití i části standardu PREMIS, jež se týká práv nakládání s dodanými soubory. Dodavatel tak například může některé události s objekty omezit či vyloučit (PREMIS Rights). Z hlediska dlouhodobé archivace se jedná o důležité informace, jež mohou ovlivnit provádění některých ochranných opatření. Využít tuto část standardu například doporučuje i Národní archiv ČR. V případě finské národní digitální knihovny je povinné popisovat události a činitele a je vytvořen lokální řízený slovník typů událostí, jež je výběrem z událostí doporučených Kongresovou knihovnou. Tyto události jsou doporučené a je povinné je zaznamenat, pokud je to u konkrétního objektu možné. Je možné dále zaznamenat i události, jež nejsou v lokálním kontrolovaném slovníku. Události jsou propojeny přes odkaz s dotčeným objektem. Pokud

však událost neodkazuje na konkrétní objekt, jedná se o událost, která se týká celého informačního balíčku, tj. jedná se o odlišný způsob zápisu události týkající se vícero objektů, než jaký využívá švédská národní knihovna. (Metadata Requirements and Preparing Content for Digital Preservation: v1.7.0)

Standard PREMIS a úkoly ve standardizaci NDK

Aktualizace verze standardu PREMIS

V uplynulém roce byl analyzován případný dopad na Standard NDK v případě, že by se ve standardu aktualizovala verze metadatového schématu PREMIS na verzi 3.0. Verze 3.0 nabízí nejen nové elementy, ale i nové funkce. Dle Angely Dappert je tato verze ve své funkci archivačních metadat robustnější. Pokud by se nevyužily nové funkce, které verze 3.0 nabízí, pak není nutné verzi PREMISu ve standardu aktualizovat, verze jsou zpětně kompatibilní (Dappert, 2018). Pouhá aktualizace verze by s sebou nesla jenom změnu názvu některých elementů a bez využití nových funkcí by to znamenalo spíše drobnou komplikaci. Dle našeho názoru má tedy smysl uvažovat o navýšení verze, pokud se při tom využijí i nové funkce standardu PREMIS.

Přechod na verzi 3 standardu PREMIS by pro Standard NDK znamenal změnu názvu dvou elementů v části PREMIS Object (původní relatedObjectIdentification a relatedEventIdentification jsou nově relatedObjectIdentifier a relatedEventIdentifier), a v části PREMIS Event by došlo k zanoření elementu eventDetail pod nově ustanovený nadřazený element eventDetailInformation. Další okamžité dopady by aktualizace standardu PREMIS ve standardu NDK neměla.

Kontrolované slovníky NDK

V dosavadní praxi NDK ale existuje několik podnětů pro úpravu Standardu NDK, jež by měly vést k větší konzistenci dat a které nesouvisejí s aktualizací verze standardu PREMIS. Jedním z nich je vytvoření kontrolovaných slovníků, a to například kontrolovaného slovníku pro elementy týkající se informací o formátech. V NDK jsou sice explicitně předepsané přijímané formáty a takové jsou do úložiště i dodávány, ale v praxi se v několika případech objevila chyba u zápisu jejich identifikátoru z registru PRONOM, kdy byl zapsán jiný identifikátor, než formát ve skutečnosti měl. Kontrolovaný slovník by měl tuto chybovost eliminovat.

Rozšíření metadatového popisu

Co stojí také za zohlednění v některé z budoucích aktualizací standardu PREMIS, je využití elementu rozšíření u objektu (objectCharacteristicsExtension) pro uložení výstupů z nástrojů pro charakterizaci (extrakci metadat), jako je tomu například v systému Archivematica. Ne vždy je totiž v metadatech možné uložit všechny informace, které o souboru zjistí charakterizační nástroje, nakolik v metadatových schématech na to neexistují vhodné elementy (např. popis dlaždic, počet vrstev kvality a dekompozičních

úrovní u formátu JPEG2000). Je tak možné například u obrazů vložit do odpovídajícího elementu PREMIS výstup z nástrojů jpylyzer, FITS, veraPDF apod.

V následujících letech bude pozornost věnována také využití standardu PREMIS pro zaznamenání signifikantních vlastností (viz výše). Průzkum zahraniční praxe ukázal, že instituce tuto možnost popisu využívají spíše jen výjimečně. Popis signifikantních vlastností je využit například v implementaci PREMIS v projektu HathiTrust, kde slouží k popisu reprezentace, pro níž byly stanoveny jako klíčové vlastnosti například počet souborů v balíčku a počet naskenovaných stránek (Elkiss, 2018; HathiTrust, PREMIS Implementation – Version 2.0, 2015). Určení signifikantních vlastnosti se považuje za nelehký úkol, jak uvádí i standard PREMIS. V ideálním případě by jejich určení mělo probíhat spolu s uživatelskou komunitou a standard PREMIS upozorňuje, že v některých případech by měl signifikantní vlastnosti a jejich hodnoty dodat vkladatel nebo kurátor archivu. Určení signifikantních vlastností tak bude zřejmě probíhat ve spolupráci s uživatelskou komunitou Standardu NDK, jímž je Formátový výbor, který je poradním orgánem Národní digitální knihovny pro oblast standardizace digitálních dat.

Dále se uvažuje pro Standard NDK využití popisu objektu na úrovni reprezentace. Události a práva se totiž mohou týkat nejen jednotlivých souborů (úroveň Object), ale i reprezentací a intelektuálních entit. Implementace popisu na úrovni reprezentace by znamenala, že objekty stejného typu nebo stejných vlastností by byly popisovány souhrnně v jedné reprezentaci, a tudíž by se události a související činitelé zaznamenávali jednou k reprezentaci, a nikoliv opakovaně ke každému souboru reprezentace. V případě několikasetstránkových monografií tento postup může znamenat snížení objemu metadat, a tedy i jejich rychlejší parsování. Je však třeba zvážit, co by tato změna znamenala pro systém LTP a dosud uložená data, a zda snížení objemu metadat by bylo dostatečným důvodem pro úpravu systému. Tato nová funkce by se mohla týkat nejen dat v balíčcích SIP, ale i v uvažovaném vnitřním metadatovém formátu úložiště NDK.

Vnitřní formát LTP úložiště

Pro uložiště Národní knihovny se dále uvažuje o vývoji jednotného vnitřního metadatového formátu, který by měl jednodušší celkovou strukturu a vycházel by z metadatového standardu PREMIS. V praxi by to znamenalo, že by se metadata z balíčků SIP s obsahem různých dat (digitalizované tištěné, zvukové, digital born) v uložišti mapovala do společného metadatového formátu, čímž by se usnadnilo a zefektivnilo získávání informací o uložených datech. Tento vnitřní formát, pokud by byl založený na standardu PREMIS, by pak zřejmě využíval i možnost popisu intelektuální entity jako objektu dlouhodobého uložení, jež je novou funkcí ve verzi 3.0. Znamenalo by to, že i intelektuální entita by se mohla popisovat elementy z části PREMIS Object a mohla by být také předmětem událostí. Bude tak například možné v metadatech zaznamenat události, jež se týkají intelektuální entity, případně celého informačního balíčku. Klasickým případem je aktualizace bibliografických metadat, jenž se jako událost aktuálně do metadat nezaznamenává. Stejně tak by bylo užitečné využít zde již výše zmíněnou úroveň

reprezentace, kdy například událost migrace formátu by se zapsala jednou pro všechny archivní kopie obrazů místo pro každý obraz zvlášť.

Pro vnitřní metadatový formát by se též vytvářely kontrolované slovníky, a to například pro výsledky událostí, jež by měly mít jasné předem dané hodnoty, aby bylo i za desítky let zřejmé, jak tedy událost dopadla. Například kontrolovaný slovník HathiTrust pracuje s hodnotami pass, warning a success a rozlišuje jimi výsledek události, jenž objekt nemění (pass) a výsledek události, která objekt mění (success) (HathiTrust, PREMIS Implementation- Version 2.0, 2015).

Také pro Standard NDK pro tvorbu balíčků SIP by měl být vytvořen kontrolovaný slovník pro výsledky událostí. Standard NDK pro plnění nabízí hodnoty, ty jsou ale pouze příkladem, nejedná se o uzavřenou množinu hodnot. V datech uložených v LTP systému Národní knihovny se tak v tomto elementu vyskytují různé hodnoty označující úspěšný průběh události a to „OK" a „successful". Ty sice obě nyní jasně indikují, že událost byla úspěšná, ale je třeba brát v úvahu skutečnost, že k takovéto podobě dat bude přistupováno i za desítky let a pak mohou být dvě hodnoty označující stejný výsledek problém.

Závěr

Metadatové schéma PREMIS je již od počátku zásadní součástí standardů NDK a jak vyplývá z výše uvedeného textu, je běžnou součástí provozů i jiných zahraničních institucí. Standard PREMIS se průběžně vyvíjí na základě nových poznatků, praxe a potřeb těchto institucí. Národní knihovna ČR tedy musí tento vývoj sledovat a průběžně analyzovat dopady nových vlastností standardu PREMIS na případnou implementaci ve Standardu NDK. Aktuální analýza využití PREMIS ve Standardu NDK poukázala na některé nedostatky v naší praxi, jenž budou v následujícím období řešeny. Zároveň standard PREMIS nabízí dosud nevyužité funkce, které mohou usnadnit správu objektů a některé dokonce zlepšit možnosti popisu pro dlouhodobé uchovávání.

Je však také možné konstatovat, že podoba, v jaké byl standard PREMIS ve standardizaci NDK implementován, je poměrně komplexní a umožňuje popsat všechny důležité vlastnosti objektů a jejich životního cyklu, což je jedním z hlavních úkolů standardu PREMIS. Případné úpravy v implementaci ve Standardu NDK znamenají náročný proces, a to nejen finančně a každá taková úprava s sebou nese riziko, není-li implementace dostatečně promyšlena. Je třeba vždy promyslet, jak úpravy ovlivní pozdější export a využitelnost dat.

———

Dedikace

Článek byl realizován v rámci institucionálního výzkumu Národní knihovny České republiky financovaného Ministerstvem kultury ČR v rámci Dlouhodobého koncepčního rozvoje výzkumné organizace.

Bibliografie

Bredenberg, Karin, Luis Faria, Miguel Ferreira, Anders Bo Nielsen, Jan Rörden, Sven Schlarb, Carl Wilson (31.5. 2019). *E-ARK Archival Information Package (AIP). Version 2.0., (DILCIS Board).* Dostupné z: https://earkaip.dilcis.eu/pdf/aip-specification.pdf

Caplan, Priscilla. (2017) *Understanding PREMIS*. 2018-11-06. Dostupné z: http://www.loc.gov/standards/premis/understanding-premis-rev2017.pdf

Cubr, Ladislav. (2017) *Autenticita a digitální informace.* (Disertační práce). Dostupné z: https://is.cuni.cz/webapps/zzp/detail/105596.

Cubr, Ladislav. (2017-2018) *Analýza standardu Premis, věnující se signifikantním vlastnostem*. Interní dokument Národní knihovny.

Dappert, Angela. (20.2. 2018) *Digital Preservation Metadata & Improvements to PREMIS in v3.0*. In: Youtube. Dostupné z https://www.youtube.com/watch?v=MU3Od6mviQs&t=3s

Elkiss, Aaron (2018). *HathiTrust PREMIS Implementation.* Dostupné z: https://www.loc.gov/standards/premis/pif/2018/iPres2018-HathiTrust.pdf

Guidelines for using PREMIS with METS for Exchange. (2017). Dostupné z: https://www.loc.gov/standards/premis/guidelines2017-premismets.pdf

HathiTrust (2015). *HathiTrust PREMIS Implementation – Version 2.0*. Dostupné z: https://docs.google.com/document/d/1UTZNIzRfelVixIYJ9nZ12tnieIe6FNZSMi9Fdhh2z5c/edit

Lavoie, Brian, Gartner, Richard. (2013) *Preservation Metadata. 2.nd edition.* 2019-05-28. Dostupné z: https://1url.cz/GMZgu

Metadata Requirements and Preparing Content for Digital Preservation: v1.7.0. National Digital Preservation Services. Dostupné z http://digitalpreservation.fi/files/Metadata-1.7.0-en.pdf

National Library of Sweden. (2013-02-04) *Creating information packages in Project Digidaily: Specification Documents*. Dostupné z: http://www.kb.se/namespace/digark/deliveryspecification/agreement/dd1/digidaily_specifications_eng.pdf

PREMIS/METS for scalability. Archivematica Wiki. 2018-02-20. Dostupné z: https://wiki.archivematica.org/PREMIS/METS_for_scalability

Premis Editorial Committee (July 2012). *PREMIS Data Dictionary for Preservation Metadata. Version 2.2*. Washington (DC): Library of Cngress. Dostupné z: https://www.loc.gov/standards/premis/v2/premis-dd-2-2.pdf

Premis Editorial Committee (June 2015). *PREMIS Data Dictionary for Preservation Metadata. Version 3.0*. Washington (DC): Library of Cngress. Dostupné z: https://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf

Using PREMIS with METS. Dostupné z: https://www.loc.gov/standards/premis/premis-mets.html

---

Poznámka o autorech

Natalie Ostráková, nar. 1983, natalie.ostrakova@nkp.cz

Pavlína Kočišová, nar. 1989, pavlina.kocisova@nkp.cz

Miroslava Beňačková, nar. 1993, miroslava.benackova@nkp.cz

Autorky v současnosti pracují v Oddělení pro standardy Národní knihovny ČR, které analyzuje, zavádí a rozšiřuje standardy potřebné pro dlouhodobou ochranu digitálních dat. Toto oddělení také zastřešuje standardizaci v rámci projektu Národní digitální knihovny (NDK), a poskytuje konzultace ostatním digitalizujícím knihovnám.