PREMIS Data Dictionary for Digital Preservation

URL: https://www.loc.gov/standards/premis/

PV070 Digitálne Knižnice

Filip Ďuračka (456172)
3.ročník B-IN UMI
27.11.2018

Predstavenie projektu

Pracovná skupina PREMIS (Preservation Metadata: Implementation Strategies) bola založená v roku 2003. Jej cieľom bol a stále je vývoj metadatového štandardu pre dlhodobé uchovávanie dát. Za týmto účelom v máji 2005 vydali prvú verziu PREMIS Data Dictionary for Digital Preservation. Tento štandard je postavený na systéme OAIS a rozširuje skorší Preservation Metadata Framework, ktorému dáva implementovateľné sémantické jednotky.

Načo je vlastne takýto štandard dobrý? Povedzme, že máme 30 rokov starý repozitár plný digitálnych dát – filmov, fotografií, hudby, kníh, vedeckých článkov… a potrebujeme sa k jednému z nich dostať. Lenže, v akom formáte je uložený? Dá sa vôbec na súčasnom hardware spustiť? Máme ešte nejaký program, ktorý by ho dokázal zobraziť? Na akom software vlastne bežal? Vzťahuje sa naň ešte copyright? Čo ak predmet pozostáva z viacerých súborov? Máme ich ešte všetky? Vieme ich poskladať naspäť do pôvodného objektu? Čo sa s objektom dialo, kým bol v repozitári? Robila sa nad ním nejaká údržba, migrovali sme ho na iný formát? Na tieto otázky a radu ďalších ponúka PREMIS implementovateľnú odpoveď, čo žiadny iný metadatový štandard zatiaľ neurobil.

Aktuálna verzia PREMISu je 3.0, vydaná v júni 2015. Štandard taktiež vyhral Digital Preservation Award v rokoch 2005 a 2012 za najväčší prínos k digitálnemu uchovávaniu dát za posledných 10 rokov.

Dátový model

PREMIS Data Dictionary je definovaný v XML a ako taký na svoju funkciu definuje 4 entity. Každá obsahuje niekoľko sémantických jednotiek. Väčšina týchto jednotiek je nepovinná, čo dáva užívateľom slobodu použiť tak komplexné / jednoduché metadata, aké potrebujú. Tie jednotky, ktoré povinné sú, definujú minimálne množstvo informácií potrebných na dlhodobé uchovávanie dát v repozitári.

V nasledujúcich odstavcoch si PREMISové entity a ich povinné sémantické jednotky popíšeme.

Entita Object

Je samotná informácia určená na digitálne uchovávanie. Každý Object si drží svoj *objectIdentifier* a jeden zo štyroch typov v jednotke *objectCategory:*

Intellectual Entity

Konkrétne dielo určené na uchovávanie (kniha, mapa, software). Môže mať viacero reprezentácií. V rámci tohto typu sa rozlišuje špeciálny typ Enviroment, ktorý reprezentuje prostredie (software aj hardware) vyžadované objektom na správnu funkcionalitu.

Representation

Objekt reprezentujúci Intelektuálnu entitu – všetky súbory, ktoré sú potrebné na úplné zobrazenie diela, napr. Naskenované strany knihy, prepis knihy v textovom súbore atď.

File

Pomenovaná a usporiadaná sekvencia bitov známa operačnému systému[1]

Bitstream

Dáta v rámci súboru s vlastnosťami relevantnými pre uchovávanie.

Všetky objekty typu File a Bitstream musia mať definovaný svoj formát v jednotke *objectCharacteristics.*

Každému objektu môže úložisko vyplniť jednotku *preservationLevel,* ktorý reprezentuje úroveň "starostlivosti" o daný objekt.

Entita Agent

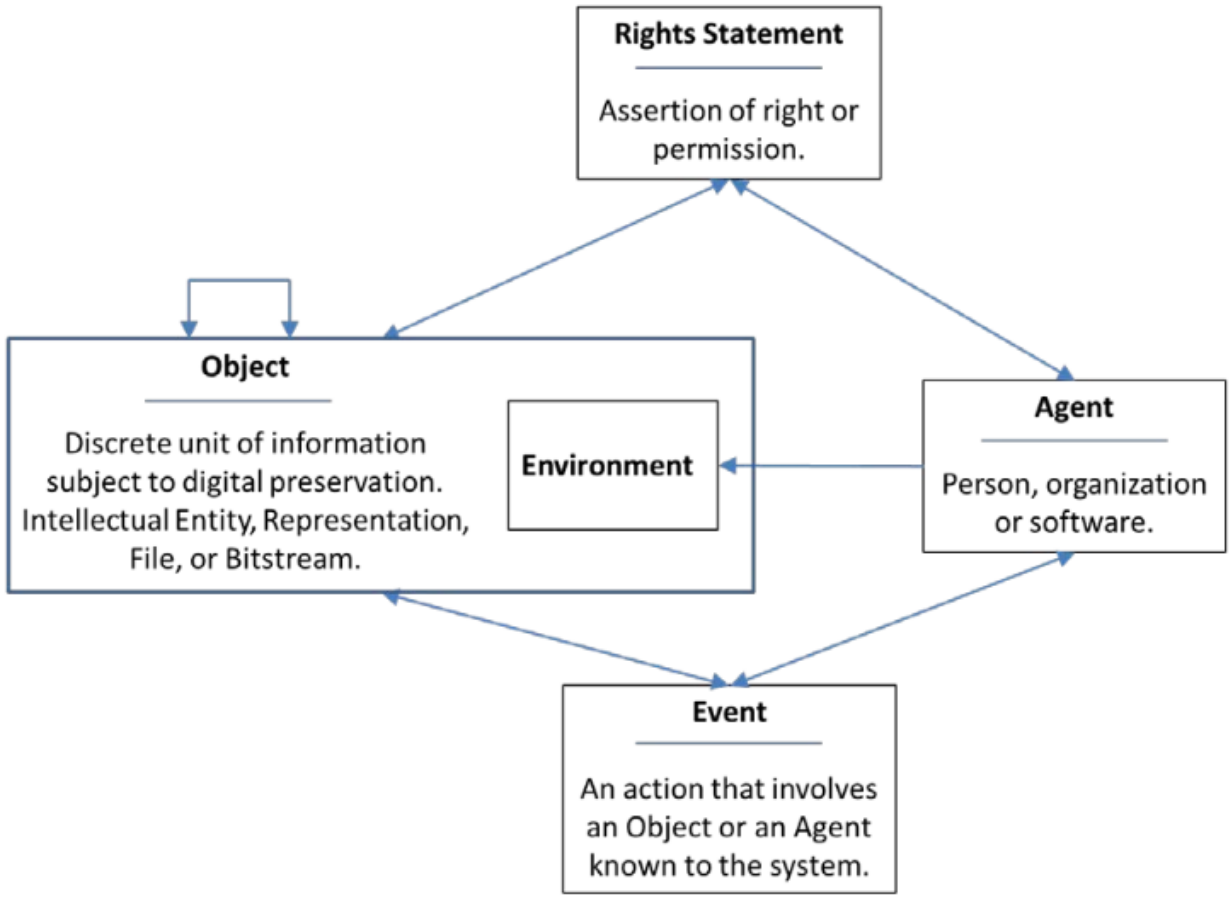
Osoba, organizácia, software či systém relevantý k udalostiam či právam viazaným k Objektu. Entita obsahuje len základné metadata o agentovi (typ, meno…), hlavné použitie má pri viazaní práv k objektom.

Entita Event

Akcia zahŕňajúca / ovlplyvňujúca aspoň jeden Objekt zahrnutý v úložisku. Všetky musia mať definovaný svoj typ a dátum a čas udalosti, a musia mať vzťah k aspoň jednému objektu. Ďalej môže obsahovať popis, čo sa s objektom stalo (*eventDetailInformation*) a ako to objekt ovplyvnilo (*eventOutcomeInformation*).

Entita Rights Statement

Súhrn práv a povolení viazaných s objektom či agentom. PREMIS si uvedomuje dôležitosť autorských práv a preto ich umožňuje týmto spôsobom v metadátach zachytiť.



Obrázok 1: Entity a možné vzťahy medzi nimi[1]

Prirodzene, pre všetky entity platí, že musia byť v rámci úložiska unikátne identifikovateľné. Toto je docielené pomocou jednotky *{entity.type}Identifier*, ktorá drží ako samotnú hodnotu identifikátora, tak aj jeho doménu (napr. ISBN). Taktiež, všetky entity môžu byť navzájom poprepájané pomocou *linking{entity}Identifier* jednotky, vytvárajúc tak logickú hierarchiu a umožnujúc zachytiť nutné vzťahy medzi jednotlivými entitami.

Medzientitné vzťahy sú zvlášť zaujímavé medzi entitami typu Object, kde PREMIS vymedzuje 3 typy:

Štrukturálny

Ide o vzťah medzi objektami tvoriacimi reprezentáciu digitálneho objektu. Nemá predsa zmysel ukladať fotky strán knihy ak ich repozitár nevie poskladať naspäť dokopy.

Derivačný

Vytvára sa pri replikácií či transformácií objektu – najčastejšie pri migrácii na iný formát.

Závislostný

Existuje hlavne na prepojenie objektov s prostredím, ktoré potrebujú na svoju správnu činnosť či správne zobrazenie.

Nie je tajomstvom že nedostatok spätnej kompatibility či prechod na úplne inú architektúru môže spraviť uchovávané dáta zbytočnými. Veď súbor, s ktorým žiadny software ani hardware nevie pracovať je zbytočná sekvencia bitov. PREMIS si je tohto plne vedomý a práve preto je entita Enviroment hodná špeciálnej pozornosti. Je to novinka najnovšej verzie štandardu a berie ohľad na dosť dôležitú súčasť uchovávania digitálnych dát, ktorú ostatné štandardy neberú do úvahy.

PREMIS je od samého začiatku implementačne nezávislý - predpokladá kooperáciu s inými metadatovými štandardmi (hlavne pre deskriptívne, technické a packaging metadata), čím konečnému používateľovi dáva slobodu voľby, ako si bude spravovať svoj repozitár. Táto nezávislosť je ďalej umocnená možnosťou rozšíriť entity Rights, Agent, a niektoré sémantické jednotky o ďalšie metadata. Tieto externé metadata môžu byť v ľubovolnom štandarde, avšak musia byť vhodne zachytené v jednotke *{entity|unit}Extension*. Aj vďaka tejto funkcionalite je PREMIS doporučovaný autormi METS na použitie s ich štandardom.

PREMIS sa taktiež veľmi spolieha na automatické generovanie metadát, teda aspoň vo valnej väčšine prípadov. Tento predpoklad je však údajne jednou z najväčších slabín tohto štandardu, pretože generovanie metadát v tomto formáte je pomerne obtiažne.

V súčasnosti je štandard PREMIS implementovaný vo viacerých archivačných repozitároch, menovite však:

- Florida Digital Archive (repozitár pre verejné univerzity a knižnice na Floride)
- Cairo - Complex Archive Ingest for Repository Objects (konzorcium vedené knižničnými službami Oxfordskej univerzity)

A k záveru praktický príklad – jednoduché metadata pre tento dokument by v PREMISe mohli vyzerať napríklad takto:

```
<premis version=”3.0”>
      <object xsi:type=”file”>
            <objectIdentifier>
                  <objectIdentifierType>MyCustomID<objectIdentifierType>
                  <objectIdentifierValue>456172-1</objectIdentifierValue>
            </objectIdentifier>
            <objectCharacteristics>
                  <format>
                        <formatName>PDF</formatName>
                  </format>
                  <creatingApplication>
                        <creatingApplicationName>
                              LibreOffice Writer
                        </creatingApplicationName>
                        <creatingApplicationVersion>
                              6.0.3.2. (x64)
                        </creatingApplicationVersion>
                  </creatingApplication>
            </objectCharacteristics>
      </object>
</premis>
```

Ako typ File, dokument musí mať okrem identifikátoru uvedený aj svoj formát. Typ *MyCustomID* by bola repozitárom dodaná špecifikácia, ako má validné ID vyzerať, 456172-1 by bola už konkrétna hodnota. Navyše môžme vidieť jeden z nepovinných údajov - dáta o aplikácií, čo súbor vytvorila.

Zhodnotenie

Vďaka mojim zberateľským tendenciám ma vždy zaujímalo, ako by sa dali archivovať digitálne produkty, či už filmy, hry, fotografie alebo hudba. PREMIS Data Dictionary v tomto ohľade pokladám za úctyhodnú iniciatívu pre dlhodobé uchovávanie takýchto digitálnych dát. Obzvlášť oceňujem pozornosť, ktorú venovali uchovávaniu informácii o prostredí potrebnom na správnu funkciu dát. Mnoho hráčov má totižto obavy, že ich obľúbené hry z detstva budú na ich staré kolená nehrateľné, lebo nebude systému, ktorý by ich spustil. Vďaka PREMISu si však potenciálny repozitár bude môcť pohodlne tieto údaje uchovať. Uchovávanie informácií o autorských právach taktiež považujem za významné plus, obzvlášť v raných fázach archivácie objektu.

Zdroje

[1] https://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf

http://www.dcc.ac.uk/resources/briefing-papers/standards-watch-papers/premis-data-dictionary

https://www.loc.gov/standards/premis/

http://www.digitalpreservation.gov/series/challenge/premis.html

http://www.dlib.org/dlib/may08/lavoie/05lavoie.html

Metadata

<dc:title>PREMIS Data Dictionary for Digital Preservation</dc:title>
<dc:creator>Filip Ďuračka</dc:creator>
<dc:subject>PREMIS Data Dictionary</dc:subject>
<dc:description>Esej popisuje a hodnotí dátový model metadatového štandardu PREMIS Data Dictionary, určeného na dlhodobé uchovávanie digitálnych dát.</dc:description>
<dc:date>2018-27-11</dc:date>
<dc:type>Text</dc:type>
<dc:format>public</dc:format>
<dc:identifier>https://www.loc.gov/standards/premis/</dc:identifier>
<dc:language>sk</dc:language>