Úvod do problematiky dlouhodobé ochrany digitálních dokumentů - díl 2.

Jan Hrabal, Zdeněk Hruška

Zatímco v prvním díle našeho seriálu jsme se věnovali důležitosti dlouhodobé ochrany digitálních dokumentů, zmínili jsme stěžejní dokument oboru (referenční model OAIS) a nastínili vhodná úložná média a strategie dlouhodobé ochrany, dnes nás čeká podrobnější pohled na problematiku formátů a metadat.

Formáty souborů

Díky formátu jsou programy schopné interpretovat data v souboru a správně je zobrazit. Určitě se vám již někdy stalo, že jste otevřeli nějaký "cizí" formát např. v MS Word a viděli jste pouze změť znaků bez jakéhokoliv smyslu a řádu. Daný program si prostě s datovým tokem (data streamem) nevěděl rady a nebyl schopný ho správě interpretovat.

Dnes není ve většině případů problém najít vhodný program na soubor, který potřebujete otevřít. Windows mají v základu programy, které si poradí s nejčastěji používanými formáty, navíc další programy jsou obecně známé a jejich stažení z internetu je otázkou okamžiku.

Pokud narazíte na neznámý formát, který není žádný aktuálně nainstalovaný software schopný zpracovat, není nic jednoduššího, než zadat příponu formátu do internetového vyhledávače a odpověď máte za pár vteřin. Většinou i s doporučením, v čem lze soubor otevřít (webové databáze souborových přípon můžete najít např. na http://www.filesuffix.com/cs/ nebo http://filext.com/).

Ale v oblasti LTP se pohybujeme v jiném časovém horizontu a na jiných úrovních - pokud nejsme schopni na našem počítači otevřít a zobrazit některé soubory, protože neznáme jejich formát, je to pro nás v podstatě jen menší nepříjemnost.

Pokud ale nebudeme v budoucnu schopni otevřít soubory z našeho repozitáře, protože jsou v neznámém formátu, bude to už pořádný problém. Mohlo by dojít k faktické ztrátě uložených informací, i když média a data budou naprosto v pořádku.

Nejčernější scénář (pokud by se nepodařilo soubory otevřít) může vést až ke ztrátě významného kulturního díla (pokud už nebude originál existovat) a přijde vniveč snaha mnoha lidí, nemluvě o promrhaných financích za několik (desítek) let. Proto je oblast, která se zabývá formáty, velmi důležitou součástí problematiky okolo LTP.

Jak jsme si již řekli v prvním díle našeho seriálu, formáty vhodné pro dlouhodobou ochranu jsou především ty, které jsou otevřené, dobře zmapované a mají širokou podporu ze strany

výrobců softwaru - tj. existuje více programů od různých výrobců, které jsou schopné soubory otevřít a zobrazit.

Pří výběru vhodných formátů pro LTP je nejvhodnější držet se doporučení komunity s ohledem na dobrou praxi a obecně přijímané standardy.

Některá kritéria výběru vhodných formátů:
zdroj: http://www.digitalpreservation.gov/formats/intro/format_eval_rel.shtml

- **udržitelnost** - jak snadné bude možné formát v budoucnu otevřít a zobrazit
    - **otevřenost formátu** - jak moc jsou jeho specifikace dostupné pro veřejnost. Jedná o zavřený formát bez zveřejněné specifikace? Nebo naopak je dokumentace volně dostupná?
    - **rozšířenost** - kolik tvůrců a uživatelů dat formát používá? Jde o běžně rozšířený kancelářský formát, nebo je omezen jen na jeden specifický hardware nebo software?
    - **transparentnost** - do jaké míry je možné analyzovat digitální reprezentaci základními nástroji (např. lidským okem za pomocí textového editoru). Transparentnost je omezena kódováním a kompresí.
    - **sebedokumentace** - sebedokumentující se digitální objekt obsahuje deskriptivní, technická, administrativní a další metadata.
    - **externí závislost** - závislost formátu na specifickém hardwaru, operačním systému, nebo reprezentačním softwaru. Přílišná závislost může v budoucnu způsobit problémy.
    - **patentová závislost** - pokud někdo vlastní patent na formát, může v budoucnu svoje právo uplatnit a používání formát bude omezeno.
    - **ochranné mechanismy** - má formát nějaké kódování? Kódování je další možná překážka v budoucích krocích dlouhodobé ochrany.
- **kvalita a funkcionalita** - schopnost formátu podporovat významné vlastnosti (significant properties - viz minulý díl našeho seriálu). U obrázků to je např. podpora vysokého rozlišení, barev, grafických efektů nebo typografie. U textových formátů to může být schopnost udržet rozvržený dokumentu, podpora fontů nebo matematických vzorců.

Vzhledem k tomu, že nežijeme v ideálním světě, každý formát splňuje výše uvedené vlastnosti jenom z části. Je na uvážení repozitáře, aby na základě svého uvážení vybral vhodné formáty.

V knihovních digitálních repozitářích se většinou můžeme setkat s relativně omezeným počtem typů dat (oproti některým jiným repozitářům, které přijímají např. nejrůznější vědecká data, multimediální formáty, nebo se zaměřují na uchovávání softwaru), proto si nyní představíme nejvhodnější formáty pro jejich uchovávání. Vycházíme při tom z ustálené praxe v paměťových institucích a také z návrhu Library of Congress (LoC)

(dostupný zde: http://www.loc.gov/preservation/resources/rfs/TOC.html).

Textová digitální díla:
Preferované formáty seřazené dle vhodnosti od nejlepšího:

1. PDF/A
2. PDF (v nejvyšší možné kvalitě, bezztrátové kompresi, s obrázky ve vysokém rozlišení)
3. Rich Text Format (RTF)
4. Otevřené formáty (např. *.odt)
5. Plain text (čistý text *.txt)
6. Široce rozšířené proprietární formáty (např. MS Word *.doc)

Digitální obrázky (rastrovaná grafika, skeny knih, map):
Preferované formáty seřazené dle vhodnosti od nejlepšího:

1. TIFF (bez komprese)
2. JPEG2000 (bezztrátový (*.jp2))
3. PNG (*.png)
4. JPEG/JFIF (*.jpg)
5. Digital Negative DNG (*.dng)
6. JPEG2000 (ztrátový (*.jp2))
7. TIFF (s kompresí)
8. BMP (*.bmp)
9. GIF (*.gif)

Audio (hudba, audioknihy, nahrávky):
Preferované formáty seřazené dle vhodnosti od nejlepšího:

1. obecně bezztrátové nekomprimované formáty - Národní archiv Austrálie doporučuje FLAC, Moravská zemská knihovna se rozhodla ukládat zdigitalizované gramodesky do formátu WAVE.

Vždy je na repozitáři, jakou formátovou politiku zvolí - je možné striktní nastavení s tím, že nebudou přijímány jiné, než vybrané formáty. V takovém případě je vhodné, aby ve spolupráci s dodavateli a uživateli dat byly stanoveny nejvhodnější formáty. Repozitář by také měl dodavatelům dat pomoci, aby pro ně bylo co nejjednodušší dodávat data ve specifikovaných formátech (např. jim doporučí vhodně softwarové nástroje na ukládání dat ve vhodných formátech).

Jinou variantou je, že repozitář bude přijímat data ve všech formátech (klidně i v neznámých, nebo exotických), ale i při největší snaze nebude schopný zaručit dlouhodobou

ochranu některých nevhodných formátů. Dodavatelé dat pak musí buď formáty změnit, nebo toto riziko přijmout.

V reálném provozu repozitáře je možné využít pro práci s formáty několika nástrojů, které výrazně pomáhají s identifikací, validací a v některých případech i extrakcí metadat ze souborů.

Jejich seznam můžete nalézt na našem webu v sekci "Užitečné odkazy".

Metadata

Metadata se v knihovnictví tradičně používají pro podrobný popis dokumentů a významnou komponentu tvoří také ve strategii dlouhodobé ochrany. V této oblasti je však třeba proces patřičně modifikovat.

Existují různé typy metadat, která se rozlišují na základě jejich funkce a záměru. Lze je rozdělit následovně:

- popisná - poskytují popisné údaje o intelektuální entitě jako údaje o původu (autor, název) a slouží pro vyhledání a zpřístupnění digitálního objektu. Mohou také poskytovat údaje o originálním (tištěném) objektu či identifikátory. Jako příklady popisných metadatových standardů uveďme MARC21, Dublin Core, MODS, VRA Core nebo EAD (Encoded Archival Description).
- strukturální - zachycují vztahy mezi dílčími digitálními objekty a jak dohromady tvoří jednu intelektuální entitu. Například kde na webové stránce se nachází obrázek (fyzická struktura) nebo jak jdou stránky a kapitoly v knize za sebou (logická struktura).
- technická - poskytují údaje o počtu souborů a jejich velikosti, o formátech souborů a jejich dalších vlastnostech (rozlišení obrázku, délka audio souboru), upřesňují údaje o hardwaru a softwaru, na nichž mohou být digitální objekty spuštěny. U technických metadat se používají schémata, jako jsou textMD (nebo audioMD, videoMD), MIX (Metadata for Images in XML Schema) či ANSI/NISO Z39.87-2006.
- administrativní - obsahují informace o vzniku entity a o zodpovědné osobě za její správu, informace o ochranných činnostech provedených na entitě, o právech k přístupu k entitě a právech vztahující se k ochranným opatřením. Administrativní metadata jsou chápána jako ochranná metadata, neboť zbývající tři typy nejsou typické pro LTP, byť jsou pro něj esenciální.

Jednotlivé typy metadat jsou vkládána do kontejnerů. Pro tuto potřebu vznikl standard METS (Metadata Encoding and Transmission Standard), což je XML schéma propojující popisná, strukturální, technická a administrativní metadata do jednoho kontejneru. Obsahuje též souborový inventář (soubory tvořící digitální objekt). Standard METS je široce rozšířený a používaný řadou světových národních knihoven.

Významným standardem pro ochranná metadata se stal PREMIS (PREservation Metadata: Implementation Strategies). Obsahuje datový model s pěti prvky: intelektuální entita (tj. spravovaný intelektuální obsah, např. kniha), objekt (informační jednotka, se kterou pracuje koncový uživatel), událost (aktivita vztahující se na objekt nebo na činitele), činitel (osoba, organizace či software) a duševní práva. Dále obsahuje datový slovník definující sémantické jednotky, které popisují entity z datového modelu. PREMIS podporuje různé možnosti zavádění metadat, nesoustředí se na žádnou konkrétní technologii, architekturu nebo strategii.

Při tvorbě metadat je třeba si upřesnit, jaká metadata potřebujeme a kde je vezmeme, jaký mají vztah na digitální objekty, repozitář a na použitou strategii dlouhodobé ochrany (migrace, emulace, …), jaké jsou funkce použitého LTP systému, jaké jsou vztahy mezi digitálními objekty či jaké entity popisujeme.

Použité zdroje (kam se také můžete podívat pro více informací):

K formátům:
http://www.digitalpreservation.gov/formats/intro/format_eval_rel.shtml

http://www.filesuffix.com/cs/

http://filext.com/

http://www.loc.gov/preservation/resources/rfs/TOC.html

http://openpreservation.org/technology/products/

http://openpreservation.org/technology/products/jpylyzer/

http://openpreservation.org/technology/products/jhove/

http://openpreservation.org/technology/products/fido/

http://fileformats.archiveteam.org/

http://www.loc.gov/standards/premis/FE_Dappert_Enders_MetadataStds_isqv22no2.pdf

K metadatům:
http://www.planets-project.eu/docs/presentations/Dappert_PreservationMetadata.pdf

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

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

CUBR, Ladislav. *Dlouhodobá ochrana digitálních dokumentů*. 1. vyd. Praha: Národní knihovna České republiky, 2010, 154 s. ISBN 9788070505885.

HUTAŘ, Jan a Marek MELICHAR. *Analýza dostupných technologií a srovnání národních strategií v oblasti dlouhodobé archivace digitálních dokumentů.* Praha, Brno, Wellington, 2013.

LAVOIE, Brian a Richard GARTNER. *Preservation Metadata (2nd edition)*. Digital Preservation Coalition 2013. Dostupné z: http://goo.gl/oJXhcq