Rosetta – nástroj pro dlouhodobé uložení digitálních objektů

Eliška Pavlásková
Multidata Praha, s.r.o.
eliska.pavlaskova@multidata.cz
INFORUM 2012: 18. konference o profesionálních informačních zdrojích
Praha, 22. -24. 5. 2012

Abstrakt

Ex Libris Rosetta umožňuje paměťovým institucím uchovat a zároveň i zpřístupnit jejich sbírky, a to dnes i v budoucnu. Rosetta je řešením pro dlouhodobou archivaci digitálních objektů a jako taková byla vyvinuta s důrazem na flexibilitu i škálovatelnost. Systém je založen na mezinárodně uznávaných standardech (OAIS, PREMIS).

Digital preservation a jeho význam

Digital preservation nebo také dlouhodobá ochrana digitálních objektů je v oblasti digitálních knihoven a archivů žhavým tématem již po několik let. Množství digitalizovaných i v digitální podobě vznikajících objektů stále narůstá. Do digitalizace se investují stále větší a větší sumy a digitálně vznikající dokumenty se stávají nedílnou součástí kulturního dědictví, vědeckého výzkumu i běžné administrativní praxe – všechny tyto oblasti vyžadují zabezpečení přístupu k objektům a jejich využitelnosti v delším časovém období. Mnoho institucí zabývajících se sběrem, správou a šířením informací v digitální formě si uvědomuje nutnost digitální objekty nejen vytvářet, ale také o ně pečovat takovým způsobem, aby s nimi uživatelé mohli pracovat i v budoucnosti. Jedná se nejen o otázku zastarání medií, na kterých jsou data uložena, ale i o problematiku zastarávání samotné formy, ve které jsou data uživatelům předkládána. V momentě, kdy si uživatel (ať už je jím kdokoli) nebude schopen digitální objekt zobrazit a zpracovat jeho obsah, stává se objekt pouze jakýmsi reliktem, který nemá praktické využití. Návrat k zobrazitelné podobě pak může být velice náročný, a to jak po technické tak po finanční stránce.

Problematiku dlouhodobé ochrany lze pracovně rozdělit do dvou hlavních oblastí či úrovní – jsou to **ochrana bitů** (tedy samotných dat) a takzvaná **logická ochrana**. Ochrana bitů je zaměřena zejména na uchování dat v nezměněné a nepoškozené podobě. Jedná se primárně o technickou disciplínu úzce provázanou se správou a návrhem hardwarových systémů úložišť. Je třeba soustředit se nejen na propracovaný systém záloh a redundanci, ale i na možnost kontroly případného poškození souborů.

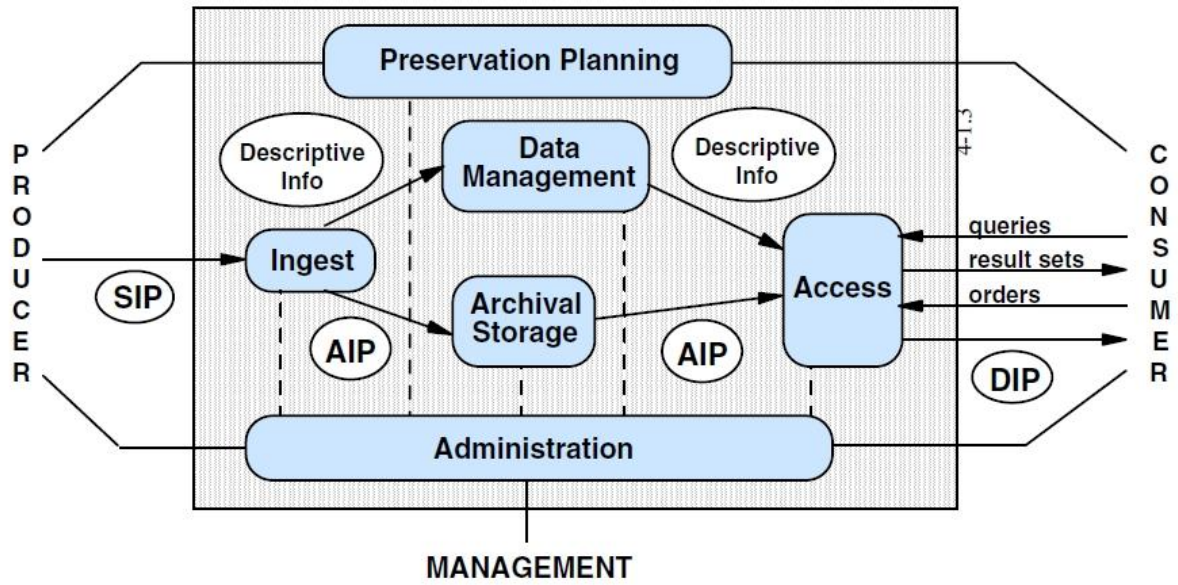
Logická ochrana je zaměřena na uchování obsahu objektů. Není zde zdaleka tak podstatné zachování přesné podoby souboru, jako spíše toho, čím je podstatný pro uživatele. V rámci logické ochrany může být například vědecký článek ve formátu doc migrován do formátu PDF/A. Jeho kontrolní součet, formát i aplikace potřebné pro zobrazení souboru se výrazně změní, nicméně obsah zůstane pro uživatele zachován v podobě, která byla institucí spravující tento objekt zvolena jako optimální. Klíčová je zde prevence zastarání objektů. Jde o soubor technik a postupů, které umožňují bezproblémové provedení tzv. ochranných akcí i podporující proces plánování a implementace těchto procesů. Systém Rosetta poskytuje modul O**chrana**, který je určen k tvorbě a následnému provedení tzv. plánu ochrany.

Obě tyto úrovně jsou samozřejmě prakticky nutnou podmínkou existence důvěryhodného digitálního repozitáře. Nicméně z hlediska technických nástrojů pro podporu obou těchto úrovní je jednoznačně lépe pokryta ochrana bitů. Logickou ochranou se na profesionální a

v praxi využitelné úrovni zabývají celosvětově v podstatě pouze dva systémy. Jak už název tohoto příspěvku napovídá – jedním z nich je systém Rosetta vytvořený společností Ex Libris.

Teoretické základy Rosetty

Jak již bylo předesláno, dlouhodobá ochrana digitálních objektů je středem zájmu odborné komunity již po delší dobu a během této doby bylo vyvinuto několik podstatných standardů, které by měly tvořit základ jakéhokoli systému s funkcemi **Long Term Preservation** (LTP). Jednoznačně zde hraje prim referenční model **Open Archival Information Systém[1]** a standard metadat pro dlouhodobou ochranu **PREMIS[3]**. Dále je třeba doporučit i používání standardních a ověřených metadatových formátů, a to jak pro popis digitálních objektů, tak pro zachycení jejich struktury. V neposlední řadě jsou zde i metadatové extraktory a jejich výstupy standardizované do formy technických metadat.
Model OAIS (viz Obrázek 1) byl původně vyvinut pro potřeby uchovávání dat z výzkumu kosmu. Rychle si však získal uznání odborné komunity a v roce 2003 se stal i ISO standardem pro dlouhodobé uchovávání digitálních objektů. Jedná se jak o teoretický rámec, tak o zdroj terminologie i o popis procesů, jež by měly probíhat v rámci důvěryhodného digitálního repozitáře. Není to tedy pouze model softwaru, ale model repozitáře jako celku. OAIS se soustředí na zachování informačního objektu jako celku – tedy nejen souboru, ale i takzvané reprezentační informace, která má zajistit dlouhodobou srozumitelnost, použitelnost a autenticitu. Tato informace by měla být součástí metadatového záznamu a spolu se souborem tvoří takzvaný archivní informační balíček (AIP). Obsah tohoto balíčku je třeba v průběhu času měnit v závislosti na potřebách tzv. cílové (designated) komunity, a to tak aby byl pro členy této komunity stále srozumitelný a použitelný. Rosetta je s modelem OAIS plně kompatibilní.



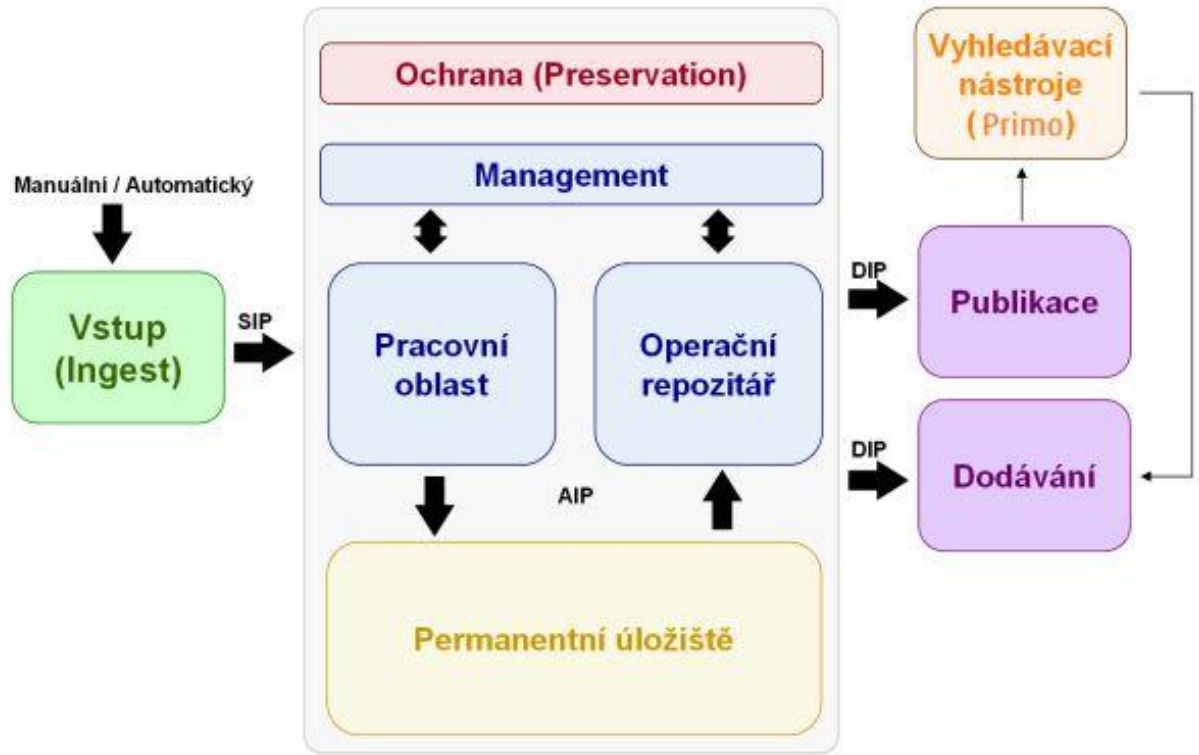
Obrázek 1 - model OAIS

Standard PREMIS se skládá ze dvou částí – datového modelu a datového slovníku. Obojí je určeno k popisu ideálního obsahu metadatového záznamu sloužícího k dlouhodobé ochraně digitálního obsahu. Datový model definuje entity, s nimiž digitální repozitář pracuje. Klíčovým pojmem je zde **intelektuální entita** – tedy smysluplná jednotka z hlediska repozitáře. Touto entitou může být například jedna konkrétní monografie. Intelektuální entita se skládá z tzv. **reprezentací** – tedy vyjádření objektu v různé formě určených k různým

účelům. Může se jednat například o archivní kopii monografie ve formátu tiff a o reprezentaci určenou k prezentaci uživatelům, která je ve formátu jp2000 (a používá například nižší rozlišení). Samotná reprezentace se skládá z jednotlivých souborů – tedy například obrázků, z nichž každý reprezentuje jednu stránku dané monografie. Tento datový model Rosetta plně přejímá a využívá tuto strukturu digitálního objektu.
Druhou částí standardu PREMIS je datový slovník, který určuje, co má být obsahem metadat pro dlouhodobou ochranu. Nejedná se o rigidní metadatový standard, ale spíše o vymezení obsahu, který by měl být v rámci metadat uchováván. Tento slovník je v Rosettě implementován ve formě interního metadatového formátu DNX.
Co se konkrétních metadatových formátů týče, Rosetta využívá několik standardů. Z hlediska popisných metadat je primárně implementován formát **Dublin Core**. Nicméně metadata v jakémkoli jiném formátu mohou být přidána jako tzv. zdrojová metadata (source metadata). Strukturní informace vážící dohromady celý digitální objekt je uchovávána na bázi formátu **METS[2].**
V rámci Rosetty jsou využívány standardní extraktory technických metadata (aktuálně je dostupný **Novozélandský metadatový extraktor[5]** a **JHOVE[4]**) a vzhledem k otevřenosti systému je možné implementovat i další nástroje dle přání zákazníka.

Vnitřní struktura Rosetty

Jak již bylo předesláno, Rosetta je plně kompatibilní s modelem OAIS, a proto mu poměrně výrazně odpovídá i její vnitřní struktura. Systém je rozdělen do několika modulů (viz Obrázek 2).
Logicky prvním modulem je **Vstup (Ingest)**, který je určen k příjmu a zpracování materiálů vstupujících do systému (dle terminologie OAIS se jedná o SIP). Objekty je možno vkládat jak manuálně, tak automatizovaně ve velkých dávkách, a to za pomoci API rozhraní tzv. vstupní aplikace.



Obrázek 2 - Moduly Rosetty

Dále se balíček SIP dostává do **Pracovní oblasti** (Working Area), kde prochází validačními procesy, obohacením (např. o technická metadata) a případnými dalšími úpravami, které mu dodají strukturu, která je archivem vyžadována pro jeho dlouhodobé uložení. Ze SIP se tím stává AIP (Archival Information Package). Pracovní oblast je doplněna **Operačním repozitářem**, který umožňuje provádění obdobných procesů nad objekty již v repozitáři uloženými. Správa objektů probíhá v rámci modulu **Management.** V administračním rozhraní je možno provádět monitoring a audit veškerých procesů, které se v systému odehrávají. Kompletní AIP je uložen do **Permanentního úložiště** (Permanent), tedy do místa, kde je digitální objekt trvale uložen. Data jsou zde uchovávána tak, aby byla minimalizována jejich závislost na jakémkoli konkrétním softwaru a na databázi. Díky tomu mohou být v případě nutnosti obnovena i bez využití systému Rosetta. V rámci AIP je uchovávána kompletní informace o digitálním objektu, a to včetně metadat popisujících historii objektu. Samozřejmostí je taká uložení verzí objektu a možnost kdykoli se k nim vrátit.
Zveřejnění a publikaci objektů má na starosti modul **Výstup.** Dle terminologie OAIS se zde již pracuje s Dissemination Information Packet (DIP). V Rosettě je výstup rozdělen do dvou částí. První z nich – publikační – má na starosti šíření metadatových záznamů (využíván je protokol OAI-PMH) v podobně, kterou mohou sklidit systémy, které chtějí se záznamy dále pracovat. V rámci produktů Ex Libris se jedná zejména o discovery systém Primo. Samotné objekty jsou posléze zobrazovány prostřednictvím prohlížečů, které jsou určeny pro různé druhy objektů. V případě potřeby zobrazení specifického objektu nebo specifického způsobu zobrazení, je možné snadno do Rosetty implementovat nový prohlížeč. Přístup k objektu je řízen na základě k tomu určených metadat.
Posledním modulem systému Rosetta je **Ochrana** (Preservation). I tento modul se skládá ze dvou základních funkčních celků – ze znalostní báze a z nástroje pro plánování preventivních ochranných akcí. Znalostní báze se obsahuje několik knihoven, z nichž nejpodstatnější je knihovna formátů, která ukládá informace o datových formátech. Knihovna je vytvořena na základě celosvětově uznávaného registru formátů PRONOM[6] a umožňuje jak ukládání znalostí na lokální úrovni, tak i jejich sdílení na úrovni globální. Na základě informací obsažených v této knihovně probíhá tzv. analýza rizik, která je dále podkladem pro tvorbu plánu ochrany. Součástí modulu je nástroj pro podporu rozhodování, který umožňuje testování jednotlivých variant plánu i následné provedení vybrané strategie.

Zákaznická komunita Rosetty

Rosetta byla vyvinuta ve spolupráci s Novozélandskou národní knihovnou, kde byla také prvně implementována v roce 2008. Samotné tvorbě systému předcházela dlouhá fáze teoretického návrhu, během které byly zohledněny jak potřeby knihovny samotné tak odborný vývoj v oblasti digital preservation.
Rosetta si však rychle získala i další zákazníky mezi knihovnami. Významná je například **Bavorská státní knihovna** nebo konsorcium německých knihoven **Goportis.** Systém však je implementován i v akademické sféře. Zde je zajímavá zejména spolupráce **se Spolkovou vysokou školou technickou v Curychu (ETH Zürich).** Zde plánují uložit do Rosetty nejen své stávající rozsáhlé digitální sbírky, ale chtějí se také soustředit na problematiku dlouhodobého uložení dat z vědeckého výzkumu (tedy nikoli vědeckých výstupů např. ve formě článků, ale přímo primárních datasetů). Z archivní sféry se uživatelem Rosetty staly **Archivy Nového Zélandu.**
Rosetta má tedy stabilní zákaznickou základnu, se kterou Ex Libris pravidelně konzultuje další vývoj systému a jeho směřování. Zákazníkům je zároveň k dispozici i systém pro sdílení znalostí založený na principu Wikipedie. Společnost Ex Libris také sleduje odborný vývoj

v oblasti digital preservation a účastní se i významných projektů. Příkladem může být projekt PrestoPrime zaměřený na dlouhodobé uchovávání audiovizuálních materiálů.

Použité zdroje

1. Consultative Committee for Space Data Systems. *Reference Model for an Open Archival Information System (OAIS)* [online]. Washington (D.C.) : CCSDS, January 2002. [1, xiii, 139 s.]. Recommendation for Space Data System Standards, CCSDS 650.0-B-1. Blue Book, Issue 1. Dostupný z WWW: <http://public.ccsds.org/publications/archive/650x0b1.pdf>.

2. LIBRARY OF CONGRESS. *METS*: *Metadata encoding and transmission standard* [online]. March 22, 2012 [cit. 2012-04-27]. Dostupné z: http://www.loc.gov/standards/mets/

3. LIBRARY OF CONGRES. *PREMIS*: *Preservation metadata maintenance activity* [online]. April 12, 2012 [cit. 2012-04-27]. Dostupné z: http://www.loc.gov/standards/premis/

4. JSTOR AND THE PRESIDENT AND FELLOWS OF HARVARD COLLEGE. *JHOVE*: *JSTOR/Harvard Object Validation Environment* [online]. Copyright 2003-2009 [cit. 2012-01-31]. Dostupné z: http://hul.harvard.edu/jhove/

5. NATIONAL LIBRARY OF NEW ZEALAND. *Metadata Extraction Tool* [online]. [cit. 2012-01-31]. Dostupné z: http://meta-extractor.sourceforge.net/

6. *PRONOM* [online]. Richmond : National Archives, 2002 [cit. 2010-04-04]. Dostupné z WWW: <http://www.nationalarchives.gov.uk/PRONOM/Default.aspx>.