Úvod do problematiky dlouhodobé ochrany digitálních dokumentů - díl 3.

Jan Hrabal, Zdeněk Hruška

Ve třetím dílu našeho seriálu si na vzorových případech ukážeme, že open source hnutí se projevuje i v oblasti dlouhodobé ochrany digitálních dokumentů a také se seznámíme s některými LTP projekty v České republice.

Historické pozadí

Projekty dlouhodobé archivace v ČR navazují na projekty orientující se na procesy digitalizace a postupně se s nimi vyvíjí. Knihovní instituce začaly s digitalizací a ukládáním digitálních dat experimentovat počátkem 90. let. Dlouhodobá ochrana se v té době zabývala pouze ochranou bitstreamu a nikoli logickou ochranou. Problematice dlouhodobé ochrany se až do roku 2005 nevěnovala přílišná pozornost. Tohoto roku byl publikován dokument Koncepce trvalého uchování knihovních sbírek tradičních a elektronických dokumentů v knihovnách ČR do roku 2010. V tomto dokumentu je již potřeba dlouhodobé ochrany zdůrazněna, je zmíněna hrozba ztráty kulturního dědictví kvůli změnám technologií a zmíněna nutnost zprovoznění LTP systému. Téhož roku vznikl v NK ČR Referát pro digitální knihovnu NFS, z něhož se posléze vyvinul Odbor digitálních fondů (ODiF). Ten v současnosti spravuje a zpřístupňuje fondy digitálních dokumentů a pokrývá oblast dlouhodobé ochrany. Skládá se ze tří stěžejních oddělení. Oddělení archivace webu má na starosti vyhledání, registraci, ochranu a zpřístupňování domácích webových zdrojů. Oddělení pro standardy provádí veškeré aktivity při zavádění standardů potřebných pro dlouhodobou ochranu a promítá je do vývoje LTP systému. Třetí oddělení se věnuje naplnění koncepčního rozvoje LTP systému.

Evropská Unie si uvědomovala nezbytnost zachování kulturního dědictví a začala projekty dlouhodobé ochrany podporovat. Jedním z nich byl DigitalPreservationEurope (DPE, 2006 - 2009), do nějž se zapojila i Národní knihovna ČR a díky tomu řešila tuto problematiku na světové úrovni se zahraničními odborníky. Účast v DPE vyústila ve spolupráci na projektech CASPAR a PLANETS.

Překlady mezinárodních norem

Značným přínosem pro českou komunitu byl překlad základních ISO norem pro dlouhodobou ochranu. ISO 14 721 (též známá jako OAIS) a ISO 16363 byly přeloženy v roce

2014 zásluhou Ladislava Cubra ve spolupráci s Úřadem pro technickou normalizaci, metrologii a státní zkušebnictví.

Velmi pozitivně je nutné také hodnotit překlad certifikační normy Data Seal of Approval, na kterém se podílel Jan Hutař, Andrea Fojtů, Marek Melichar a Eliška Pavlásková. Překlad je volně dostupný na webu Karlovy univerzity: http://dsa.cuni.cz

Oblasti dlouhodobé ochrany se v ČR nevěnují jen knihovny, ale též archiváři v čele s Národním archivem, který do svého workflow implementoval jako jednu z částí systém Archivematica.

Se stejným systémem experimentuje projekt LTP Pilot, který realizuje ÚVT Masarykovy univerzity ve spolupráci s Moravskou zemskou knihovnou. Projekt je financován z Fondu rozvoje CESNET a jeho délka je 14 měsíců (začal 1. září 2014) a měl by prověřit možnosti systému hlavně v následujících oblastech:

- možnosti integrace s prostředím úložiště CESNETu
- ověření nároků Archivematiky na HW, infrastrukturu, provoz na virtuálních serverech
- zálohování a obnova dat
- testování systému – zátěžové testy, prostupnost dat, možnosti přenosu konfigurace z jedné instalace na jinou
- možnosti propojení s dalšími systémy (DSpace, systémy dodávající data)
- rozšiřitelnost systému o další nástroje (např. metadata extraktor či editor)
- různé scénáře pro ingest dat – velké objemy dat, exotická data, různé formáty
- simulace událostí a vytváření reportů
- tvorba dokumentace v češtině
- posuzování Archivematiky z hlediska norem ISO 16363 a OAIS

Pokud vše proběhne podle plánu, bude projekt velkým přínosem jak pro CESNET, tak i menší instituce, pro něž by Archivematika (jakožto open source software) mohla být zajímavým řešením a alternativou ke komerčním LTP systémům.

Open source LTP řešení

Archivematica

Archivematica je systém pro dlouhodobou ochranu digitálních dat, který je vyvíjen kanadskou firmou Artefactual Systems jako open source (zdrojový kód i dokumentace je dostupná pod AGPL3 a Creative Commons licencí). Na vývoji se podílí i další instituce - UNESCO, City of Vancouver Archives, Mezinárodní měnový fond a další významné knihovny a univerzity z celého světa. Samotný systém je zdarma, peníze na vývoj proudí od institucí, které si zaplatí instalaci, integraci s dalšími systémy, za školení nebo vývoj dalších nástrojů a funkcionalit, které jim schází.

Archivematica běží na Linuxu a to konkrétně na Ubuntu 12.04 LTS, ale v podstatě je možné ji nainstalovat na jakýkoliv počítač, který odpovídá hardwarovým požadavkům (nejsou velké, i několik let staré notebooky by měly splňovat minimální konfiguraci, která je ovšem opravdu jen minimální konfigurací a systém by měl sloužit jen k základnímu otestovaní, na běžný provoz je potřeba větší výkon, protože Archivematica je docela "žrout" výkonu.

Archivematica používá mikroslužby (microservices), což jsou v podstatě programy, které vykonávají jednotlivé funkce příjmu a zpracování dat. Některé jsou produkty třetích stran (většina z nich je pod GNU GPL licencí), nebo si uživatelé mohou naprogramovat vlastní, pokud jim ty původní nevyhovují.

AV je vystavěna na OAIS (ISO 14721) modelu , takže se zde setkáte se SIP, AIP i DIP balíčky, tak jak jsou definovány v OAIS. Samozřejmostí je používání PREMIS, METS, Dublin Core a dalších mezinárodních metadatových formátů.

Základní strategií uchování je normalizace formátů - AV používá pro uchovávání pouze formáty s otevřenými standardy, i když konkrétní volbu nechává na správcích systému a je tu tak prostor pro flexibilitu. Svým přístupem podporuje i další strategie - uchovává totiž původní data, takže podle potřeby je možné provádět jejich emulaci, ale zároveň má implementovaný identifikátor a databázi formátů a jejich rizik, čímž podporuje možnost formátové migrace.

Systém se může pochlubit přehledným GUI v podobě rozhraní webového prohlížeče, které logicky provází zpracováním jednotlivých balíků dat. Co by mohlo být zajímavé pro akademické knihovny, tak je podporován automatický příjem z DSpace za pomoci OAI-PMH a OAI API. Archivematica umí také pracovat s různými druhy úložných médií - RAID pole, magnetické pásky nebo jiná média, i když zde narážíme na její slabou stránku a tou je, že v systému zatím chybí nástroje na správu a uložení více kopií na různých místech. Nicméně i tak se jedná o zajímavý SW, který je možné provozovat samostatně jako LTP řešení, nebo může být jen jedním kolečkem v systému a zajišťovat např. jen ingest a normalizaci dat. V tomto ohledu je možná velká flexibilita a záleží pouze na instituci, jak se rozhodne Archivematiku využít.

Více informací o Archivematice je možné nalézt na webu: www.archivematica.org.

RODA

Systém RODA je úplný digitální repozitář poskytující funkcionality všech hlavních jednotek referenčního modelu OAIS. Je schopen ingestu, řízení a umožňuje přístup k digitálnímu obsahu. Roda je budována v jazyce JAVA a je tak nezávislá na jakékoli architektuře, designu, hardwaru či strategii uchovávání. Nicméně jsou v něm zabudovány aktivity pro strategii migrace. Podporuje stovky formátů a díky jejímu nastavení je schopna se v budoucnu rozšířit o další nové formáty a o nástroje pro lepší podporu uchovávání. Jistota kvality a použití metadat zajišťuje autenticitu záznamů a umožňuje tak stopovat všechny změny, které proběhly na digitálním záznamu. Všechny akce provedeny v repozitáři jsou z důvodu bezpečnosti a odpovědnosti ukládány do logu. RODA se skládá z modulů jako jsou Ingest, který validuje vložené informace a také extrahuje technická metadata z ingestovaných souborů; Data management spravující deskriptivní metadata založené na Mezinárodním standardu pro archivní popis (International

Standard for Archival Description, ISADg) a podporované standardem EAD/XML; modul Archivace je pod správou Fedora Commons, pro jednodušší zálohovatelnost jsou data ukládána do systému odděleně od metadat. Modul přístupu, přičemž přístup je zajištěn webovým prohlížečem. Modul administrace obsahuje pokročilé administrativní rysy (správa uživatelů, reporty, konfigurace ingest workflow, prohlížeč logů, správa oprávnění). Dále se tam nachází proces provádění plánování akcí pro uchovávání, jenž je zabudován v modulu administrace.