

Archivematica – vybrané zahraniční projekty

Archivematica – selected foreign projects

Zdeněk Hruška / Moravská zemská knihovna v Brně (The Moravian Land Library), Kounicova 996/65a, 601 87 Brno

Resumé: Článek přináší přehled vybraných zahraničních digitálních repozitářů, které využívají open source systém pro dlouhodobou archivaci Archivematica, a představuje tak možnosti využití tohoto systému v digitálním workflow jednotlivých institucí.

Klíčová slova: Archivematica, digitální repozitář, zahraniční praxe, dlouhodobá ochrana digitálních dat, open source

Summary: The paper brings a survey of some selected foreign digital repositories making use of the open source system for long-term archiving Archivematica, introducing the options for applying this system in the digital workflow of various institutions.

Keywords: Archivematica, digital repository, foreign practice, long-term preservation, digital preservation, digital archiving, open source

Tento příspěvek vznikl v rámci řešení výzkumného projektu Fondu rozvoje CESNET č. 516R1/2014 s názvem „Pilotní projekt pro low-barrier přístup k ochraně digitálního obsahu (LTP-pilot)“.

1 Úvod

Archivematica, open source systém pro dlouhodobou ochranu digitálních dat, si za dobu své krátké existence získala poměrně značnou popularitu v mnoha paměťových institucích po celém světě. Za tímto úspěchem je několik klíčových vlastností systému Archivematica. Jednou z těchto vlastností je variabilita při využití systému – je možné s ním absolvovat celý postup od příjmu dat po vytvoření balíčků AIP a DIP,¹ nebo jej použít jen jako jednu dílčí část v širším systému – např. pro normalizaci vstupních dat do formátů vhodných pro dlouhodobé uložení. Pro uživatelskou komunitu je neméně důležité to, že systém je vyvíjen ve shodě s normou OAIS² a jako nízkonákladové řešení

¹ Každý informační balíček má svoje místo v příjmu, uložení a výstupu dat v digitálním repozitáři. Balíčky jsou definovány modelem OAIS (viz poznámka č. 2) a jedná se o tři typy balíčků: SIP (Submission Information Package – vstupní informační balíček), AIP (Archival Information Package – archivní informační balíček) a DIP (Dissemination Information Package – výstupní informační balíček).

² Referenční model OAIS (Open archival information system – Otevřený archivační informační systém) je jedním ze základních stavebních kamenů oboru digitální archivace dat. Obsahuje termíny a slovník pojmů, hlavně ale stanovuje základní entity a procesy v digitálním repozitáři. Byl přijat jako norma ISO 14 721, která byla v roce 2014 přeložena do češtiny jako ČSN ISO 14721. Historický vývoj a kontext vzniku OAIS, jeho komponenty a koncepty jsou zpracovány v článku od Christophera A. Leeho: http://rc.statearchivists.org/Handler.ashx?Item_ID=984F92BA-1474-4659-A6E6-D1A23B01A2BB. Pro zájemce o OAIS je k dispozici revidovaný úvod od Briana Lavoie v angličtině: <http://dx.doi.org/10.7207/TWR14-02>. Původní anglická verze OAIS je dostupná na webu: <http://public.ccsds.org/publications/archive/650x0m2.pdf>.

na bázi open source. Pro mnohé paměťové instituce a jejich digitální repozitáře je tato varianta lákavým doplňkem.

Další výhodou je modularita systému a možnost úprav či výměn jednotlivých mikroslužeb³ za jiné, které více vyhovují potřebám daného repozitáře. Pokud má organizace k dispozici programátory, může jít i cestou vývoje vlastních mikroslužeb.

Díky neustálému vývoji, vylepšování a široké komunitě uživatelů, aktivně podporované vývojáři, se Archivemata do budoucna jeví jako zajímavá alternativa k ostatním systémům pro dlouhodobou ochranu/archivaci digitálních dat (LTP – Long Term Digital Preservation), což potvrzuje i celosvětové rozšíření tohoto systému. Vzhledem k tomu, že jeho tvůrci (společnost Artefactual Systems⁴) jsou z Kanady, největší rozšíření a dlouhodobější projekty využívající systém Archivemata můžeme najít právě na území severní Ameriky, postupně se ale tento systém rozšiřuje i do dalších částí světa. Testování a implementace systému probíhá také v Evropě, především ve Velké Británii a Německu, ale i v České republice – kromě projektu LTP-pilot pracuje se systémem Archivemata i Národní archiv ČR a počítá s ním i systém ProArc⁵ – ten by měl být ve své archivační části schopen napojení na Archivemata. V tomto článku představíme vybrané projekty ze zahraničí, které reprezentují možnosti implementace systému Archivemata do postupů a procesů jednotlivých institucí.

2 City of Vancouver Archives

Archiv města Vancouveru byl spolu s Archivem Mezinárodního měnového fondu vůbec prvním uživatelem systému Archivemata a od začátku také intenzivně spolupracuje se společností Artefactual Systems na testování, hodnocení a dalším rozvoji systému.

Ze své pozice městského archivu organizace přijímá a spravuje velké množství dat městské správy i soukromých institucí,⁶ dosavadním vrcholem byl závazek k uložení dat z olympijských a paralympijských her konaných ve Vancouveru v roce 2010. Na rozdíl od běžně přijímaných dat, u nichž existuje určitá možnost ovlivnit dodávané formáty, data z olympiády a paralympiády obsahovala velké množství nejrůznějších formátů, v celkovém objemu přes 20 TB.⁷

Archivemata slouží v archivu jako přijímací nástroj, jednotlivé mikroslužby provádějí testy na malware⁸, kontrolují integritu, extrahují metadata, identifikují, charakterizují a normalizují formáty. Výsledné balíčky AIP a DIP jsou pak postoupeny dále – balíček

³ Mikroslužby (microservices) jsou nástroje nebo programy, které jsou součástí systému Archivemata a vykonávají jednotlivé funkce příjmu a zpracování dat – např. antivirová kontrola, přiřazení identifikátoru, validace formátů a další. Tyto mikroslužby jsou snadno nahraditelné a pokud některá z nich uživateli nevyhovuje, může použít jiný dostupný nástroj a do Archivematicy ho implementovat.

⁴ <https://www.artefactual.com/>

⁵ NEZBEDOVÁ, Martina. ProArc – open source řešení pro produkci a archivaci digitálních dokumentů. In: *INFORUM 2015: 21. ročník konference o profesionálních informačních zdrojích, Praha 26.–27. května 2015* [online]. Praha: Albertina icome Praha, 2015. ISSN 1801-2213. Dostupné také z: <http://www.inforum.cz/pdf/2015/nezbedova-martina.pdf>.

⁶ CITY OF VANCOUVER ARCHIVES. *City of Vancouver Archives* [online]. c2015 [cit. 2015-03-08]. Dostupné z: <http://vancouver.ca/your-government/city-of-vancouver-archives.aspx>.

⁷ DINGWALL, Glenn. *Building a Digital Archives at the City of Vancouver* [online]. 2010 [cit. 5. 3. 2015]. Dostupné z: http://www.interpares.org/display_file.cfm?doc=ip3_canada_dissemination_cs_dingwall_ip3-isym03_2010.pdf.

⁸ Obecně škodlivý software, jako jsou viry, trojské koně, spyware a další.

AIP je uložen do archivního uložště, balíček DIP je přesunut do systému AtoM⁹, kde k němu mají přístup uživatelé.

Formátová politika archivu je nastavena tak, že pokud je to možné, jsou data normalizována do formátů vhodných pro dlouhodobou ochranu (např. z formátu MS Word jsou převedena do PDF). Pokud to možné není (jde o exotické formáty nebo formáty, které nemají obdobu vhodnou pro dlouhodobou ochranu), jsou data přijímána a ukládána v původním formátu a archiv se je i tak bude snažit dlouhodobě uchovat.¹⁰

City of Vancouver Archives je vzorovým uživatelem systému Archivematica – tento systém zde prošel mohutným testováním a je používán v nepozměněné podobě, bez úprav ze strany archivu. Organizace je navíc aktivně zapojena do uživatelské komunity a dalšího vývoje systému (díky tomu může promítat svoje potřeby do dalších verzí tohoto systému).

3 Simon Fraser University Archives

Archiv Univerzity Simona Frasera v kanadském Burnaby má několikaleté zkušenosti se systémem Archivematica, který tvoří páteří systém digitální ochrany na univerzitě.

Od roku 2011 probíhal pilotní projekt testování systému Archivematica¹¹ s důrazem na zpracování a následné nahrání balíčků DIP do nástroje AtoM. Testovaná data se skládala především z kancelářských dokumentů, obrázků a digitalizovaného audia ve formátu WAV a AIFF.¹²

Po testech následovala doporučení k dalšímu rozvoji systému Archivematica – především v oblasti uživatelského rozhraní, workflow, dávkového zpracování (batch processing), zpracování chyb (error handling), správy a systémové administrace.

Další projekt archivu se zabýval uchováváním elektronické pošty z e-mailového klienta Zimbra.¹³ Byla vybrána neaktivní e-mailová schránka, která byla využívána přes 10 let – za tu dobu se zde nashromáždilo na 10 000 e-mailů. Vzhledem k tomu, že je Zimbra proprietární software¹⁴, bylo nutné provést migraci do formátu vhodnějšího pro dlouhodobou archivaci. Ve spolupráci s Artefactual Systems byl vyvinut nástroj na konverzi formátu Zimbra do Maildir¹⁵, který uchovává jednotlivé zprávy. To je výhoda, pokud je nutné e-maily rozdělit na ty, jež je možné zpřístupnit veřejně, a na ty, které obsahují osobní nebo obchodní informace a jejichž dostupnost je kvůli tomu nutné omezit. Tento

⁹ AtoM je nástroj od stejné společnosti jako Archivematica, který může mimo jiné sloužit jako prezentační vrstva pro přístup k uloženým datům. Více na <http://www.accesstomemory.org/>.

¹⁰ MUMMA, Courtney, Glenn DINGWALL a Sue BIGELOW. A First Look at the Acquisition and Appraisal of the 2010 Olympic and Paralympic Winter Games Fonds: or, SELECT * FROM VANOC_Records AS Archives WHERE Value="true";. *Archivaria: The Journal of Association of Canadian Archivists*. 2011, č. 72. Dostupné z: <http://journals.sfu.ca/archivar/index.php/archivaria/article/view/13361/14666>.

¹¹ Nejříve ve verzi 0.7.1, v roce 2012 proběhl update na verzi 0.8, která byla používána po zbytek projektu.

¹² ARTEFACTUAL SYSTEMS. *Simon Fraser University Archives Digital Preservation Strategy: Project report* [online]. 2012 [cit. 5. 3. 2015]. Dostupné z: https://wiki.sfu.ca/departments/archives/images/9/95/ProofConceptReport_20120418.pdf.

¹³ <http://www.zimbra.com/>

¹⁴ Takový software, který nemá otevřený zdrojový kód, uživatel jej nemůže měnit a jeho používání je typicky upraveno licenčními podmínkami. Jedná se např. o operační systém Windows, programy kancelářského balíčku MS Office, aj.

¹⁵ <http://cs.wikipedia.org/wiki/Maildir>

problém byl vyřešen pomocí open source nástroje Muse, umožňujícího tagování e-mailů. Zároveň byly přílohy e-mailů převedeny na otevřené formáty vhodné pro dlouhodobé uložení. Bylo rozhodnuto, že balíček AIP bude obsahovat soubory ve formátu Maildir a balíček DIP ve formátu mbox, který je snadněji zobrazitelný a je tedy vhodnější pro zpřístupnění koncovým uživatelům.¹⁶

Oba projekty měly za úkol zjistit, jak Archivematica vyhovuje potřebám a budoucím plánům archivu Univerzity Simona Frasera. Bylo přihlédnuto k tomu, že Archivematica se nadále vyvíjí. Jedním z výstupů byla i doporučení a požadavky na další rozvoj a archiv intenzivně spolupracoval se společností Artefactual Systems.

Na předešlé zkušenosti navázaly práce na vytvoření repozitáře pro závěrečné práce studentů univerzity,¹⁷ kde je Archivematica částí celkového workflow a je zodpovědná pouze za vytváření archivních balíčků, kdežto ostatní procesy zajišťují jiné služby (např. fyzické uložení a několikanásobné záložní kopie zajišťuje systém LOCKSS¹⁸).

Jako další projekt vzniká na univerzitě repozitář vědeckých dat¹⁹ jako komplexní řešení, které má uživatelům nabízet více funkcí, z čehož digitální ochrana je jen jednou z mnoha. Repozitář jako takový je postaven na softwaru Islandora, dlouhodobé uložení zajišťuje Archivematica spolu s úložnými systémy – opět zde platí princip vícenásobných záloh – jak v síti univerzity, tak v systému LOCKSS.

Jednotlivá oddělení univerzity a další poskytovatelé dat mohou data dodávat na offline médiích (externích nebo optických discích), nebo prostřednictvím sítě. Tato data jsou shromažďována a připravena k příjmu (ingest) – použití konkrétní pipeline²⁰ záleží na tom, jaká data jsou přijímána – pro textové dokumenty a obrázky je použita jiná pipeline, než pro velké soubory (např. audio a video), která je vybavena větším výkonem.²¹ Balíčky AIP jsou přesunuty do úložiště AIP, která je v současnosti zálohována na různých místech v rámci univerzity; vzdálené redundantní zálohy jsou teprve ve fázi plánování.²²

V repozitáři jsou tři instalace systému AtoM – první slouží k testování, druhá pro veřejný přístup a třetí pro práci s archiváliemi, importy a balíčky DIP. Do budoucna je plánována čtvrtá instalace pro dokumenty, které nemohou být volně zpřístupněny přes internet – nacházet se bude na terminálu ve studovně archivu a bude sloužit k prezenčnímu studiu materiálů. Rozhodnutí o rozdělení repozitáře na čtyři nezávislé instalace bylo provedeno kvůli výkonu, aby případný vysoký počet přihlášených uživatelů negativně neovlivňoval výkon třetí instalace systému AtoM, která je určená pro práci archivářů.

¹⁶ ARTEFACTUAL SYSTEMS. *Simon Fraser University Archives Digital Preservation Strategy: Project report* [online]. 2012 [cit. 5. 3. 2015]. Dostupné z: https://wiki.sfu.ca/departments/archives/images/9/95/ProofConceptReport_20120418.pdf.

¹⁷ JORDAN, Mark. *Preservation of Digital Theses at SFU* [online]. 2012 [cit. 6. 3. 2015]. Dostupné z: <http://summit.sfu.ca/system/files/iritems1/10884/Preservation%20of%20Digital%20Theses%20-%20Jordan.pdf>.

¹⁸ <http://www.lockss.org/>

¹⁹ <http://www.lib.sfu.ca/data-repository>

²⁰ Pipeline se dá do češtiny v oboru informačních technologií přeložit jako „tok dat“. V kontextu tohoto článku pipeline znamená instalaci systému Archivematica, kterou proudí zpracovávaná data. Více pipelines znamená více instalací systému Archivematica v digitálním repozitáři. Tak mohou např. dva pracovníci provádět ingest, aniž by byl systém nadměrně zatížen (což by nastalo, kdyby byla pouze jedna instalace tohoto systému).

²¹ Schéma architektury digitálního repozitáře v Archivu Univerzity Simona Frasera je dostupná zde: <https://wiki.sfu.ca/departments/archives/images/3/31/ArchivematicaArchitecture.pdf>.

²² SFU ARCHIVES AND RECORDS MANAGEMENT DEPARTMENT. *DIGITAL REPOSITORY PROJECT: Repository Infrastructure Requirements* [online]. 2014 [cit. 8. 3. 2015]. Dostupné z: <https://wiki.sfu.ca/departments/archives/images/e/e5/RepositoryInfrastructureRequirements.pdf>.

Jako vylepšení základního vyhledávání v systému Archivematica vyvinuli v archivu Univerzity Simona Frasera vlastní dotazovací nástroj AIP Query Tool, který vytváří statistiky z balíčků AIP (např. druhy a počty formátů nebo velikost a počty uložených objektů).

Jedním z výsledků testovacích projektů univerzitního archivu je i analýza potřeb dalšího rozvoje systémů Archivematica a AtoM. Zajímavé je, že v této analýze, která obsahuje 14 bodů (odstupňovaných podle priority), se systému Archivematica týká pouze šest z nich, a to hlavně v částech se střední a nižší prioritou, zbytek je zaměřen na úpravy a vylepšení AtoMu – tedy identifikované nedostatky se týkají spíše zpřístupnění dat. Jediný problém s vysokou prioritou u systému Archivematica se týká příjmu e-mailů, kdy testovaná verze neuspěla při příjmu e-mailů ve formátu maildir, což byl preferovaný formát pilotního projektu.

4 University of British Columbia Library

Knihovna Univerzity Britské Kolumbie úzce spolupracuje se společností Artefactual Systems od roku 2011 – byla provedena diferenční analýza²³, naplánovány pilotní projekty, proběhly první testy systému Archivematica, vznikly procesy pro tzv. digital-born dokumenty²⁴, byl zahájen projekt archivace webu a testovala se integrace se systémem CONTENTdm²⁵ a s digitální knihovnou DSpace.²⁶

Archivematica zde slouží pro příjem balíčků SIP, DIP jsou pak předávány do ICA-AtoM a CONTENTdm, které slouží jako prezentační vrstva.²⁷ Balíčky AIP jsou uloženy a zálohovány jak na lokálním úložišti, tak v síti LOCKSS²⁸.

Institucionální repozitář Univerzity Britské Kolumbie, označovaný cIRcle, je postavený na DSpace a je rovněž propojený s Archivematicou. Jednotlivé dokumenty nebo celé kolekce lze z DSpace exportovat jako balíčky AIP, které si převezme Archivematica, přijme je jako své balíčky SIP a pak s nimi dále pracuje. Je zde opět několik instalací Archivematica (pipelines), každá je nastavena na příjem jiného druhu dat. Např. pipeline Wolf je vysoce automatizovaná a je určená pro příjem videopřednášek, které jsou vkládány do cIRcle. Cílem je co největší integrace pipelines (např. propojení s webovými stránkami univerzity – Digital Collections, RBSC Archives a cIRcle), aby byly zajištěny co nejplynulejší postupy a digitální ochrana se stala automatickou součástí správy digitálních dat na UBC.²⁹

²³ Gap analysis – tedy zjištění rozdílu mezi reálným a požadovaným stavem.

²⁴ Jedná se o dokumenty, které vznikly rovnou jako digitální a nikdy neměly papírovou verzi – např. elektronické články, e-knihy, aj.

²⁵ www.contentdm.org

²⁶ ARTEFACTUAL SYSTEMS. *University of British Columbia Library Persistent Digital Collections Implementation Plan: Final project report Summary version*. 2012. Dostupné z: <http://diginit.sites.olt.ubc.ca/files/2012/07/UBCLibrary-PersistentDigitalCollectionsPlan-ProjectReportCondensed-1.pdf>.

²⁷ SPROUT, Bronwen a Sarah ROMKEY. *UBC Library's Digital Preservation Strategy* [online]. 2014 [cit. 9. 3. 2015]. Dostupné z: https://circle.ubc.ca/bitstream/handle/2429/47021/Sprout_B_et_al_UBC_Library_Digital.pdf?sequence=1.

²⁸ LOCKSS (Lots of Copies Keep Stuff Safe – hodně kopií udržuje věci bezpečné) je projekt zahájený Standfordskou univerzitou. Jedná se o síť akademických repozitářů, které si navzájem zálohují svá data a udržují jejich několikanásobné kopie. Tak jsou data zajištěna proti ztrátě nebo zničení. Více na webu projektu: <http://www.lockss.org/>.

²⁹ How We Digitize: Digital Preservation 2. In: WILL, Evan. *UBC LIBRARY BLOGGERS* [online]. 2015 [cit. 2015-03-09]. Dostupné z: <http://digitize.library.ubc.ca/digitizers-blog/how-we-digitize-digital-preservation-2/>.

5 Columbia University

V letech 2011–2013 probíhalo na Kolumbijské univerzitě zpracování Archivu Fordovy nadace – Programu mezinárodních stáží (Archives of the Ford Foundation International Fellowships Program). Cílem projektu bylo archivovat a dlouhodobě uchovat dokumenty a další digitální soubory z toho programu.³⁰ Celkově se jednalo se o 3,6 TB dat a 350 000 souborů. Zároveň se projekt musel vypořádat s několika výzvami:

1. 245 různých formátů souborů
 - a. kancelářské dokumenty
 - b. audio a video materiály
 - c. databáze
 - d. e-mailová korespondence
 - e. webové stránky
 - f. průzkumy, rozhovory, statistické zprávy
 - g. kolekce dat a další
2. dlouhé názvy souborů a souborových cest (více než 260 znaků)
3. 10 jazyků a 7 nerománských znakových sad (včetně arabštiny a indštiny)
4. nedostatek popisných metadat
5. omezení přístupu k některým datům, která se dělila na tři typy
 - a. veřejně dostupná data
 - b. interně dostupná data
 - c. data nedostupná až do roku 2075

Před samotným příjmem dat bylo nutné provést mnohé úpravy – většina souborů byla převedena na formáty vhodné pro dlouhodobé uložení (např. mbox pro e-maily, XML pro MS Access), případně byla extrahována data ze zip a rar³¹ archivů a rozdělena do jednotlivých balíčků SIP. Archivemata zde byla využita k příjmu balíčků SIP, použity byly standardní funkce, které systém nabízí – přiřazení UUID³² jednotlivým balíčků, virová kontrola, normalizace názvů souborů (viz výše uvedený velký počet jazyků a znakových sad), formátová identifikace, extrakce metadat a generování souborů METS.³³

Balíčky AIP byly posléze uloženy do dlouhodobého úložiště (jedna záloha se nachází on-site, druhá off-site a třetí v páskové knihovně ve spřátelené instituci), balíčky DIP byly nahrány do Fedory, která slouží jako prezentační vrstva pro přístup k datům.

³⁰ SOKOLOVA, Dina a Jane GORJEVSKY. *Adding Metadata and Ingesting Large Born-Digital Archives with Archivemata*. 2014. Dostupné z: http://academiccommons.columbia.edu/download/fedora_content/download/ac:180721/CONTENT/AddingMetadata_SAA_Forum_2014.pdf.

³¹ Rar je souborový formát pro kompresi dat podobný formátu zip. Na rozdíl od něj nabízí některé pokročilejší funkce, ale nejedná se o otevřený formát (nejsou volně k dispozici jeho zdrojové kódy).

³² UUID (universally unique identifier – univerzální unikátní identifikátor) je standardní identifikátor, který se používá v repozitářích a digitálních knihovnách pro identifikaci každého uloženého objektu. Skládá se z 36 znaků (32 číslic a písmen a 4 pomlček) a díky systému výpočtu je zajištěno, že žádné přidělené UUID nebude duplicitní (šance na duplicitu se prakticky rovná nule) a to bez nutnosti používat centrální autoritu (jako např. v případě ISBN), proto jednotlivá UUID může přidělovat svým objektům každý repozitář zvlášť.

³³ SOKOLOVA, Dina a Jane GORJEVSKY. *Infrastructure Development: Multiple Digital Content Types in a Single Collection*. 2014. Dostupné z: https://library.columbia.edu/content/dam/librarywebsecure/behind_the_scenes/ford-ifp/MultipleContentTypes_DP_2014.pdf.

6 Council of Prairie and Pacific University Libraries – COPPUL

COPPUL³⁴ je sdružení několika desítek univerzitních knihoven v západní Kanadě, které využívají svého uskupení pro výměnu zkušeností, ale i pro vyjednávání s dodavateli softwaru, elektronických databází a dalších služeb, aby byla zajištěna co nejvýhodnější cena pro jednotlivé knihovny.

Ve spolupráci se společností Artefactual Systems nyní COPPUL svým členům nabízí systém Archivematica *jako službu* (jdou tedy cestou nazývanou Digital Preservation as a Service – DPaaS). Sdružení zajišťuje propagaci, vstup nových členů a financování vstupních nákladů. Artefactual System má na starosti správu účtů a serverů, instalaci a technickou podporu uživatelů. Jednotlivé zapojené instituce přispívají poskytnutím serverů a úložným prostorem.³⁵

Výhodou uvedeného postupu je snížení výdajů na nákup nebo tvorbu vlastního LTP řešení, na technickou podporu a správu systémů, nezanedbatelným přínosem je i vytvoření společné komunity. Hosting a uložení digitálních objektů jsou vyřešeny v rámci partnerství jednotlivých institucí COPPUL, a to s využitím cloudu. Přístup je pak možný díky webovému portálu, důraz je kladen na plnění právních podmínek Kanady a Britské Kolumbie.

V této chvíli jsou do projektu zapojeny tyto univerzity:

- MacEwan University,
- Mt Royal University,
- University of Lethbridge,
- University of Saskatchewan,
- University of Victoria,
- Vancouver Island University.

7 Konrad-Zuse-Zentrum für Informationstechnik Berlin (The Zuse Institute Berlin)

V berlínském Centru Konráda Zuseho pro informační technologie (zkráceně Zuseho institut Berlin, ZIB) v současné době probíhá fáze budování a testování systému LTP, který se skládá z řady open source nástrojů a ve kterém Archivematica slouží pro příjem dat (ingest).

V ZIB pracují s různými druhy dat od partnerských institucí. Archivematica byla upravena tak, aby vyhovovala jejich potřebám – hlavně z důvodu nutnosti doplňování nejrozličnějších metadat do METS v balíčku AIP.

Vznikající repozitář očekává příjem různorodého materiálu, proto bude nutné udělat hodně manuální práce s daty před samotným příjmem. Zpracované AIP a vytvořené DIP pak poputují dále a na další správu dat bude využíván systém iRods³⁶. Balíčky AIP jsou uloženy na pásky a balíčky DIP importovány do repozitáře, který je postaven na systému Islandora, a data jsou pak dostupná uživatelům podle potřeby. Administraci dat je možné provádět přes propojení Islandory a iRods, přičemž balíčky AIP zůstávají uloženy na páskách, pouze DIP jsou kvůli přístupnosti uloženy online.

³⁴ <http://www.coppul.ca>

³⁵ Archivematica. COPPUL. COPPUL: Council of Prairie and Pacific University Libraries [online]. 2014 [cit. 2015-03-10]. Dostupné z: <http://www.coppul.ca/programs/archivematica>.

³⁶ <http://irods.org/>

Zuse Institute Berlin sponzoruje³⁷ naprogramování funkce re-ingestu balíčku AIP, aby bylo možné upravit a změnit metadata, ale zachovat UUID a existující metadata PREMIS (tato funkce by měla být dostupná v Archivematica ve verzi 1.5³⁸). V budoucnu by také měla být dostupná renormalizace³⁹ binárního obsahu pro migrační účely (pokud by se v instituci změnila strategie uchovávání).⁴⁰

8 ArchivesDirect

COPPUL není jediná organizace, která nabízí systém Archivematica jako DPaaS. Ve spolupráci společnosti Artefactual Systems a DuraSpace⁴¹ se od října 2014 do února 2015 uskutečňoval pilotní projekt Archivematicy v cloudu, do kterého byly zapojeny vybrané univerzity v USA.⁴² Pro veřejné používání byla služba zpřístupněna v březnu 2015, pod názvem ArchivesDirect.⁴³

DuraCloud⁴⁴ je služba, která umožňuje správu dat a jejich kopií – po nahrání dat je možné zvolit počet kopií i jejich zálohování do cloudových úložišť od různých poskytovatelů⁴⁵. DuraCloud rovněž zajišťuje kontrolu dat na úrovni bitů, nenabízí však ochranu na logické úrovni, jedná se tedy spíš o pokročilý zálohovací systém.

Archivematica nabízí funkce příjmu dat, obohacování o metadata a normalizaci jako strategii dlouhodobé ochrany. Nedostatečně ale pokrývá oblast správy dat na úložišti, neřídí počet ani umístění kopií, nekontroluje jejich integritu. Spojení obou systémů se tedy jeví jako cesta správným směrem – kombinace silných stránek by měla vytvořit zajímavé řešení, které eliminuje většinu slabín jednotlivých samostatných částí.

9 Závěrem

Jak je z výše uvedených příkladů patrné, využití systému Archivematica v institucích po celém světě je rozmanité a jeho vlastnosti umožňují široké zapojení do postupů a funkcí digitálního repozitáře. Od vydání beta verze v září 2012 prošel systém značným vývojem a podstoupil řadu testovacích projektů i ostré nasazení v provozu velkých archivů či knihoven, kde si našel své místo vedle dalších open source nástrojů pro digitální archivaci.

V budoucnu se se systémem Archivematica budeme pravděpodobně setkávat více i v našem prostředí – využít ho jako základ pro vlastní LTP řešení plánuje např. projekt

³⁷ Je možné si u společnosti Artefactual Systems zaplatit vývoj takové funkce systému Archivematica, která je pro repozitář důležitá, ale v současnosti ji Archivematica „neumí“. Vývojáři systému mají tak zajištěno financování a komunita zase rozvoj systému, protože všechny nové funkce jsou stále dostupné jako open source.

³⁸ Archivematica 1.5.0. *Archivematica wiki* [online]. 2015 [cit. 2015-08-25]. Dostupné z: https://wiki.archivematica.org/Development_roadmap:_Archivematica#Archivematica_1.5.0.

³⁹ Re-normalizace je myšleno převedení souborů buď na nové verze starších formátů, nebo přímo nové formáty, které bude v budoucnu repozitář používat.

⁴⁰ Shepherding the bits. THE ZUSE INSTITUTE BERLIN. ZIB [online]. c2014 [cit. 2015-03-10]. Dostupné z: <http://www.zib.de/features/shepherding-bits>.
<http://duraspace.org/>

⁴² The Archivematica + DuraCloud “Soup-to-Nuts” Preservation Service Launches a Beta Test. DURASPACE. *DuraSpace.org* [online]. 2014 [cit. 2015-03-17]. Dostupné z: <http://www.duraspace.org/node/2314>.

⁴³ <http://archivesdirect.org/>

⁴⁴ <http://duracloud.org/>

⁴⁵ Odpadá tedy riziko závislosti pouze na jednom poskytovateli.

ARCLib, který ve spolupráci s Národní knihovnou ČR, Moravskou zemskou knihovnou, Masarykovou univerzitou a odborníky z dalších institucí připravuje Knihovna Akademie věd ČR.⁴⁶ Návrh projektu byl podán v květnu letošního roku (2015) do programu NAKI II, k vyhlášení vybraných projektů by mělo dojít na konci tohoto roku; pokud bude projekt podpořen, můžeme se dočkat vývoje českého řešení LTP, které bude postaveno na technologiích open source a kde bude i Archivematica hrát významnou roli.

Použité zdroje:

Archivematica 1.5.0. *Archivematica wiki* [online]. 2015 [cit. 2015-08-25]. Dostupné z: https://wiki.archivematica.org/Development_roadmap:_Archivematica#Archivematica_1.5.0.

Archivematica. COPPUL. COPPUL: *Council of Prairie and Pacific University Libraries* [online]. 2014 [cit. 2015-03-10]. Dostupné z: <http://www.coppul.ca/programs/archivematica>.

ARTEFACTUAL SYSTEMS. *Simon Fraser University Archives Digital Preservation Strategy: Project report* [online]. 2012 [cit. 2015-07-13]. Dostupné z: https://wiki.sfu.ca/departments/archives/images/9/95/ProofConceptReport_20120418.pdf.

ARTEFACTUAL SYSTEMS. *University of British Columbia Library Persistent Digital Collections Implementation Plan: Final project report Summary version*. 2012. Dostupné z: <http://diginet.sites.olt.ubc.ca/files/2012/07/UBCLibrary-PersistentDigitalCollectionsPlan-ProjectReportCondensed-1.pdf>.

BARTOŠEK, Miroslav. *Dlouhodobá ochrana digitálních dat a systém Archivematica* [online]. Prezentace na semináři Otevřené repozitáře 2015, VUT Brno 12. – 13. 5. 2015 [cit. 2015-08-20]. Dostupné z: <http://slideplayer.cz/slide/5251412/>.

CITY OF VANCOUVER ARCHIVES. *City of Vancouver Archives* [online]. c2015 [cit. 2015-03-08]. Dostupné z: <http://vancouver.ca/your-government/city-of-vancouver-archives.aspx>.

DINGWALL, Glenn. *Building a Digital Archives at the City of Vancouver* [online]. 2010 [cit. 5. 3. 2015]. Dostupné z: http://www.interpares.org/display_file.cfm?doc=ip3_canada_dissemination_cs_dingwall_ip3-isym03_2010.pdf.

⁴⁶ BARTOŠEK, Miroslav. *Dlouhodobá ochrana digitálních dat a systém Archivematica* [online]. 2015 [cit. 2015-08-20]. Dostupné z: <http://slideplayer.cz/slide/5251412/>.

How We Digitize: Digital Preservation 2. In: WILL, Evan. *UBC LIBRARY BLOGGERS* [online]. 2015 [cit. 2015-03-09].

Dostupné z: <http://digitize.library.ubc.ca/digitizers-blog/how-we-digitize-digital-preservation-2/>.

JORDAN, Mark. *Preservation of Digital Theses at SFU* [online]. 2012 [cit. 2015-06-03]. Dostupné z: <http://summit.sfu.ca/system/files/iritems1/10884/Preservation%20of%20Digital%20Theses%20-%20Jordan.pdf>.

MUMMA, Courtney, Glenn DINGWALL a Sue BIGELOW. A First Look at the Acquisition and Appraisal of the 2010 Olympic and Paralympic Winter Games Fonds: or, SELECT * FROM VANOC_Records AS Archives WHERE Value="true";. *Archivaria: The Journal of Association of Canadian Archivists*. 2011, č. 72.

Dostupné z: <http://journals.sfu.ca/archivar/index.php/archivaria/article/view/13361/14666>.

NEZBEDOVÁ, Martina. ProArc – open source řešení pro produkci a archivaci digitálních dokumentů. In: *INFORUM 2015: 21. ročník konference o profesionálních informačních zdrojích, Praha 26. – 27. května 2015* [online]. Praha: Albertina icome Praha, 2015. ISSN 1801-2213. Dostupné také z: <http://www.inforum.cz/pdf/2015/nezbedova-martina.pdf>.

SFU ARCHIVES AND RECORDS MANAGEMENT DEPARTMENT. *DIGITAL REPOSITORY PROJECT: Repository Infrastructure Requirements* [online]. 2014 [cit. 2015-08-03]. Dostupné z: <https://wiki.sfu.ca/departments/archives/images/e/e5/RepositoryInfrastructureRequirements.pdf>.

Shepherding the bits. THE ZUSE INSTITUTE BERLIN. *ZIB* [online]. c2014 [cit. 2015-03-10]. Dostupné z: <http://www.zib.de/features/shepherding-bits>.

SIMON FRASER UNIVERSITY ARCHIVES. *Archivematica Architecture* [online]. 2014 [cit. 2015-08-25].

Dostupné z: <https://wiki.sfu.ca/departments/archives/images/3/31/ArchivematicaArchitecture.pdf>.

SOKOLOVA, Dina a Jane GORJEVSKY. *Adding Metadata and Ingesting Large Born-Digital Archives with Archivematica*. 2014. Dostupné z: http://academiccommons.columbia.edu/download/fedora_content/download/ac:180721/CONTENT/AddingMetadata_SAA_Forum_2014.pdf.

SOKOLOVA, Dina a Jane GORJEVSKY. *Infrastructure Development: Multiple Digital Content Types in a Single Collection*. 2014. Dostupné z: https://library.columbia.edu/content/dam/library-websecure/behind_the_scenes/ford-ifp/MultipleContentTypes_DP_2014.pdf.

SPROUT, Bronwen a Sarah ROMKEY. *UBC Library's Digital Preservation Strategy* [online]. 2014 [cit. 9. 3. 2015]. Dostupné z:

http://elk.library.ubc.ca/bitstream/handle/2429/47021/Sprout_B_et_al_UBC_Library_Digital.pdf.

The Archivematica + DuraCloud "Soup-to-Nuts" Preservation Service Launches a Beta Test. DURASPACE. *DuraSpace.org* [online]. 2014 [cit. 2015-03-17].

Dostupné z: <http://www.duraspace.org/node/2314>.