Archivematica – open source systém pro digitální archivaci

RNDr. Miroslav Bartošek, CSc.
Masarykova univerzita, Ústav výpočetní techniky

Tento text 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]

Abstrakt: Článek poskytuje základní přehledovou informaci o systému Archivematica. Tento systém je určen pro podporu dlouhodobé archivace digitálních informací v souladu s referenčním modelem OAIS. Je vyvíjen kanadskou společností Artefactual Systems a poskytován jako bezplatný open source software. Rostoucí počet institucí využívá nebo plánuje využít systém Archivematica jako nízkonákladové řešení, které umožní řešit palčivé problémy dlouhodobého uchování digitálních dat bez nutnosti masivních finančních investic do drahých komerčních produktů.

Klíčová slova: Dlouhodobé uchování digitální informace; digitální archivace; Archivematica; referenční model OAIS; nízkonákladová řešení.

1. Úvod

Problematika dlouhodobého uchování digitálních dat (LTP – Long-Term Preservation nebo DP – Digital Preservation) byla až donedávna výhradní doménou velkých institucí typu národních knihoven či národních archivů, které disponovaly potřebnými mandáty, financemi a expertními zdroji. Tyto instituce se často zaměřily na budování velkých monolitických systémů postavených na nákladnějších komerčních řešeních (příkladem je systém Rosetta od firmy ExLibris). Avšak pokroky v oblasti teorie a praxe spolu s rostoucími potřebami zabývat se digitální archivací i v menších institucích vedly k poznání, že i s omezenými zdroji lze začít vytvářet vlastní řešení – a není nutné čekat, až co nabídnou velcí hráči. Jedním ze systémů, které se objevily v posledních letech a podporují tento trend, je systém Archivematica.

1.1 Představení systému

Archivematica[2] je volně dostupný open source systém na podporu dlouhodobé archivace digitálních informací. Systém je vyvíjen kanadskou společností Artefactual Systems Inc. ve spolupráci s akademickými a paměťovými institucemi od roku 2008. Impulsem pro vznik systému byla (a) poptávka po nízkonákladovém řešení[3] dlouhodobého uchování digitálních informací, (b) dostupnost velké škály open source nástrojů pro podporu digitální archivace,

[1] Cílem projektu LTP-pilot bylo prověřit možnosti, nároky a omezení systému Archivematica; ověřit jeho použitelnost pro logickou dlouhodobou archivaci vybraných typů dokumentů a sbírek; vytvořit základní dokumentaci systému pro systémové administrátory a kurátory digitálních dat.

[2] http://www.archivematica.org

[3] Příkladem nízkonákladového přístupu je třeba projekt POWRR – Preserving Digital Objects With Restricted Resources viz http://commons.lib.niu.edu/handle/10843/13610 .

kterým však chybělo propojení do uceleného systému použitelného širší komunitou digitálních kurátorů. Deklarovaným cílem systému Archivematica je poskytnout archivářům a knihovníkům s omezenými technickými a finančními kapacitami nástroje, metodologii a sebedůvěru k tomu, aby mohli sami začít s archivací jejich digitálních informací.

Prototyp systému vznikl s cílem ověřit pracovní hypotézu[4], že použitelný a bezplatný LTP systém je možné vytvořit mapováním dostupných open source nástrojů na jednotlivé procesy funkčního modelu OAIS[5]. Systém byl zpočátku vyvíjen na zakázku Archivu města Vancouver a Archivu Mezinárodního měnového fondu, posléze se zapojily další instituce a širší komunita uživatelů. Beta-verze systému byla uvolněna počátkem roku 2009, první produkční verze o dva roky později. Poslední verzí zveřejněnou k datu odevzdání tohoto článku je verze 1.4 z května 2015.

Archivematica je systém, který integruje sadu volně dostupných nástrojů pro komplexní zpracování digitálních objektů od jejich příjmu a vložení do archivu až po zpřístupnění uživatelům dle modelu OAIS a dalších standardů a doporučení, a to s využitím specifických formátově orientovaných ochranných postupů. Pro implementaci funkcí digitální archivace používá Archivematica technologii tzv. mikroslužeb (micro-services): každá mikroslužba představuje jeden dílčí krok při postupném zpracování informace určené k uchování a je obvykle realizována zvolenými nástroji. Mikroslužby jsou řetězeny do pracovních postupů reprezentujících příslušné funkce modelu OAIS. Celý systém je uživatelem řízen a monitorován pomocí webové aplikace. Tím, že lze jednoduše měnit pracovní postupy a používané nástroje (lze je např. postupně nahrazovat novými dokonalejšími nástroji, jakmile budou k dispozici), může systém pružně reagovat na změny jak technologií pro tvorbu digitálních informací, tak i technologií pro jejich správu a uchování.

1.2 Rizika pro digitální informaci a digitální ochrana

Na rozdíl od tištěných dokumentů, které dokáží přežít a předat autentickou informaci po velmi dlouhou dobu bez potřeby jakéhokoliv zásahu, jsou digitální informace vystaveny řadě rizik, která kriticky ohrožují jejich dlouhodobou dostupnost a využitelnost. Mezi hlavní rizika patří omezená životnost záznamových médií a složitost digitálních objektů, ale především technologický pokrok, v jehož důsledku technologie potřebné pro zpřístupnění a využití původní digitální informace rychle zastarávají a stávají se nedostupnými. Bez aktivních systematických opatření nelze u digitálních dokumentů zajistit jejich skutečně dlouhodobou dostupnost, použitelnost, integritu a autentičnost.

Digitální archivace (digital preservation) zahrnuje v principu dvě různé doplňující se složky: fyzickou ochranu, tj. potřebu uchovat původní digitální soubor (sadu bitů, proto též bit-level-preservation), a logickou ochranu, tj. potřebu uchovat schopnost přečíst a porozumět informaci obsažené v digitálním záznamu.

Fyzická ochrana se týká ochrany před ztrátou samotného digitálního záznamu nebo jeho části (samovolná degradace nosiče, úmyslné či neúmyslné smazání či změna zapsaných údajů,

[4] VAN GARDEREN, Peter a Courtney C. MUMMA. Realizing the Archivematica vision: delivering a comprehensive and free OAIS implementation. In: iPRES2013: proceedings of the 10th International Conference on Preservation of Digital Objects, 3-5 September 2013, Lisbon, Portugal

[5] Open Archival Information System (OAIS) je referenční model pro dlouhodobý digitální archiv vytvořený jako doporučení mezinárodního fóra kosmických agentur Consultative Committee for Space Date System v roce 1999 a standardizovaný v roce 2002 jako mezinárodní norma ISO-14721:2003. V roce 2012 vyšla aktualizovaná verze ISO-14721:2012 (český překlad této normy byl vydán v roce 2014 pod označením ČSN ISO 14721). Velmi kvalifikovaný čtivý přehled a zhodnocení OAIS od Briana Lavoie lze najít v [6].

ztráta či destrukce nosiče – např. při havárii nebo katastrofě). Fyzická ochrana se zajišťuje zejména vytvářením vícenásobných kopií souboru uložených v různých geograficky vzdálených úložištích a pravidelnou kontrolou jejich neporušenosti.

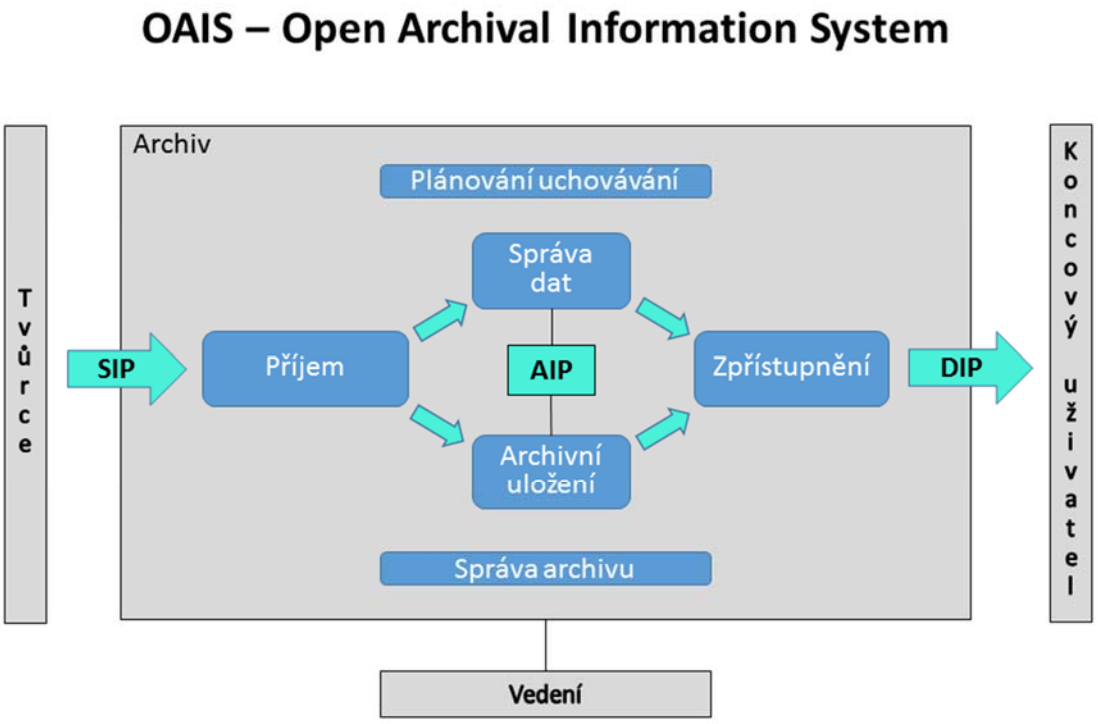
Logická ochrana se týká ochrany před neschopností uchovaný záznam přečíst (např. v důsledku budoucích technologických změn: není již k dispozici zařízení pro přečtení nosiče, neexistuje již funkční program pro dekódování formátu, není již operační systém či hardwarová platforma, na které by šlo daný program spustit), anebo porozumět jeho informačnímu obsahu (ztratily se souvislostí, kontext aj.). Logická ochrana se zajišťuje zejména tím, že spolu s uchovávaným digitálním objektem se uchovává také co nejvíce co nejpřesnějších doprovodných informací – metadat, a v průběhu času bude také nutno provádět ochranné akce, které zajistí čitelnost původní informace (jako je migrace do nového formátu, emulace původního výpočetního prostředí aj). Při logické ochraně může docházet ke změně sady bitů ve prospěch uchování čitelnosti a srozumitelnosti informačního obsahu.

Archivematica se zaměřuje na podporu procesů pro logickou ochranu (uchování informačního obsahu, jeho čitelnosti a porozumění).

1.3 Funkční a informační model OAIS

Výchozím standardem používaným dnes pro implementaci systémů dlouhodobé archivace je referenční model OAIS – Open Archival Information Standard (norma ISO 14721). Digitální obsah, který je třeba dlouhodobě uchovat, je dle modelu OAIS[6] předán tvůrcem dat správci archivu v podobě vstupního balíčku SIP (Submission Information Package). Ten je po vložení do archivu transformován do balíčku AIP (Archival Information Package), jenž je uložen do bezpečného fyzického úložiště a dále je archivem dlouhodobě spravován na základě stanovených ochranných postupů a strategií. Koncovým uživatelům je příslušný digitální obsah zpřístupněn prostřednictvím balíčku DIP (Dissemination Information Package). Viz základní schéma na obrázku 1.

[6] Standard OAIS zahrnuje tři navazující modely: *model prostředí* (interakce archivu s jeho okolím), *funkční model* (hlavní funkce archivu) a *informační model* (struktura dat a doprovodných informací v archivu). Všechny entity, vztahy a procesy jsou ve standardu OAIS detailně rozpracovány. Zde používáme velmi zjednodušený popis, který je však pro účely tohoto textu postačující.



Obr. 1 Základní schéma referenčního modelu OAIS

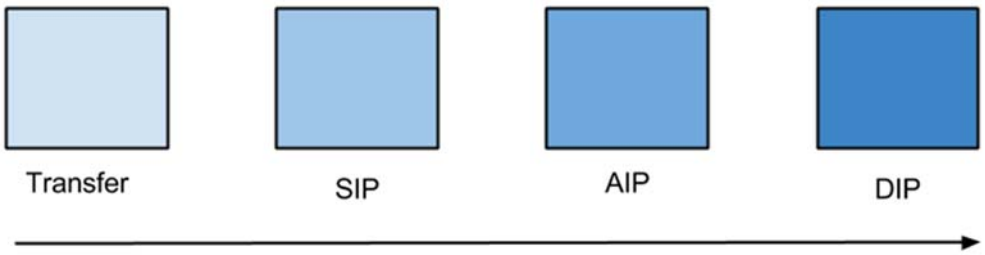
Funkční model OAIS zahrnuje šest funkčních celků: *příjem* (Ingest, přijetí informací od tvůrce a jejich příprava pro vložení do archivu), *archivní uložení* (Archival Storage, dlouhodobé uložení a ochrana informací), *správa dat* (Data Management, databáze popisných metadat archivovaných objektů spolu s administrativními daty pro fungování archivu a vyhledávání), *plánování uchovávání* (Preservation Planning, udržování archivovaných objektů s ohledem na permanentně probíhající změny externího prostředí a technologií), *správa archivu* (Administration, koordinace a provoz archivu), *zpřístupnění* (Access, zpřístupnění archivovaných objektů koncovým uživatelům).

2. Implementace modelu OAIS v systému Archivematica

Obecný funkční model OAIS byl tvůrci systému Archivematica konkretizován do podoby uživatelských scénářů, na jejichž základě byly vytvořeny konkrétní pracovní postupy implementované v systému Archivematica. Digitální informace prochází během zpracování řadou transformací, při nichž je původní digitální obsah postupně modifikován a doplňován (původní nezměněný obsah je vždy uchováván také).

Hlavní funkcí a cílem systému Archivematica je zpracovat vstupní digitální data (v terminologii systému Archivematica se nazývají *transfer*) do podoby balíčků SIP připravených pro vložení do archivu a na ně následně aplikovat formátově-orientované postupy k vytvoření archivačních balíčků AIP určených pro dlouhodobé uložení. Souběžně s vytvořením archivačních balíčků AIP lze zadat také vygenerování přístupových balíčků

DIP. Archivematica se zaměřuje primárně na vytvoření co nejkvalitnějších balíčků AIP[7]. Co se s nimi děje dále, to již Archivematica příliš neřeší, spoléhá na využití jiných externích systémů. Například ke zpřístupnění balíčků DIP uživatelům lze využít buď komplementárně vyvíjený a volně nabízený systém AtoM nebo integraci s již existujícími externími systémy dané organizace (například institucionálním repozitářem).



Obr. 2 Informační balíčky při zpracování dat v systému Archivematica[8].

2.1 Transfer

Pojem „transfer“[9] je používán ve dvou různých významech: na jedné straně označuje množinu vstupních dat a metadat (souborů a adresářů), která mají být archivována v systému Archivematica. Vedle toho ale také znamená proces předcházející službě příjem (pre-ingest), při němž je ze vstupních dat vytvářen balíček SIP.

Příprava a průběh procesu transfer závisí na typu digitálního obsahu a postupech zavedených v dané konkrétní instituci. Typicky může zahrnovat roztřídění vstupních souborů do adresářů, vytvoření metadat popisujících tyto soubory a připojení vstupní informace, která musí být uchována spolu s vlastním digitálním obsahem – jako jsou například licenční podmínky, dárcovské smlouvy apod. Archivematica má předdefinováno několik variant organizace vstupních dat pro transfer, současně ale umožňuje i vlastní uspořádání (vytvořit vlastní strukturu dat pro transfer).

Sérií postupných kroků (mikroslužeb), při nichž probíhá rozbalení zabalených souborů, normalizace jmen, antivirová kontrola, ověření/generování kontrolních součtů, přiřazení jednoznačných identifikátorů, identifikace formátů, extrakce metadat aj. vzniká ze vstupních dat balíček SIP, který je připraven pro vložení do archivu. Z jedněch vstupních dat může být při transferu vygenerován jeden nebo více balíčků SIP a naopak – jeden SIP muže zahrnovat i více než jednu sadu vstupních dat. Systém podporuje také funkci „backlog“ (nevyřízená práce), která umožňuje odložit rozpracované neúplné či jinak nedořešené transfery k pozdějšímu zpracování (častý postup v archivářských komunitách).

[7] Balíčky AIP jsou klíčovým informačním objektem pro LTP. Na kvalitě a úplnosti údajů v nich obsažených závisí dlouhodobé uchování a využitelnost původního obsahu. Každý balíček AIP zahrnuje nejen vlastní informaci, která je předmětem dlouhodobé archivace (tzv. Content Data Object), ale také celou řadu dalších nezbytných „podpůrných informací“: informace potřebné pro porozumění a prezentaci objektu v budoucnosti (jak na úrovni strukturální, tak sémantické), metadata pro podporu a dokumentování ochranných procesů – identifikaci, uchování kontextu, historii vzniku a změn objektu, prokázání integrity a autenticity, přístupové údaje a další. Struktura a obsah AIP jsou na obecné úrovni podrobně specifikovány v informačním modelu standardu OAIS. Konkrétní implementace AIP v systému Archivematica je popsána v systémové a uživatelské dokumentaci.

[8] Převzato z https://github.com/mjordan/archivematicaworkshop

[9] Tento koncept není v referenčním modelu OAIS definován, Archivematica ho zavádí jako doplněk na základě praktických zkušeností a potřeb uživatelů.

2.2 Příjem (Ingest)

Během příjmu dat do archivu procházejí balíčky SIP dalším zpracováním – například doplněním požadovaných metadat (popisná metadata v Dubline Core, archivační metadata v PREMIS, výstupy procesů validace, identifikace formátů atd.), optickým rozpoznáním znaků aj., zejména však také provedením *normalizace*, kdy na základě identifikace vstupního formátu dat může být digitální obsah konvertován do vhodnějšího archivního formátu. Současně může být generován i přístupový formát. Původní verze digitálních objektů je vždy uchována spolu s normalizovanou verzí. Na normalizaci navazuje další zpracování zahrnující vytvoření podrobné vstupní dokumentace, integrace všech metadat a doprovodných informací do formátu METS (viz 3.2), indexace atd. Archivematica nabízí několik různých předem definovaných postupů pro příjem v závislosti na typu či formě dat a úplnosti popisu vkládaného digitálního obsahu. Administrátor může ovšem tyto postupy libovolně upravit nebo definovat vlastní.

Proces příjmu je završen vytvořením archivačního balíčku AIP, případně též přístupového balíčku DIP, a jejich uložením do příslušného archivačního úložiště, resp. přístupového systému. Data, metadata a veškeré doprovodné informace tvořící balíček AIP jsou uloženy v adresářové struktuře ve formě souborů a jsou zapouzdřeny do jednoho balíčku podle standardu BagIt (viz 3.2).

2.3 Archivní uložení (Archival Storage)

Archivematica ukládá všechny informace a informační balíčky (transfer, SIP, AiP, DIP) ve formě souborů do souborového systému[10]. K zajištění nezávislosti na konkrétním fyzickém uložení dat využívá samostatnou komponentu Storage Service, která poskytuje logické rozhraní na vlastní archivní úložiště. Administrátor může nakonfigurovat Storage Service tak, aby se data ukládala v úložišti podle potřeby organizace. Úložištěm může být lokální nebo vzdálený souborový systém (např. NFS), sítová úložiště typu LOCKSS apod. Pro různé typy dat lze nakonfigurovat v rámci jednoho systému současně i více různých úložišť. Archivematica samotná neřeší bitovou ochranu, tj. zálohování, vícenásobné kopie, kontroly neporušenosti, obnovu po nepředvídatelných či katastrofických událostech apod. – přenechává toto na starosti úložišti.

Všechny balíčky AIP jsou v archivním uložení indexovány (využívá se vyhledávací server ElasticSearch), takže je lze v omezené míře vyhledávat a stahovat, a to jak na úrovni celých balíčků nebo jednotlivých objektů v nich obsažených, tak případně i na úrovni AIC (Archival Information Collection – informační jednotky sdružující soubor logicky spolu souvisejících balíčků AIP). V odůvodněných případech je možné balíčky AIP z archivního uložení řízeným postupem odstranit (není ale možné odstranit jednotlivé soubory z balíčku AIP).

2.4 Plánování uchovávání (Preservation Planning)

Archivematica používá dvoucestnou ochrannou strategii – *normalizaci* během příjmu a *zachování* původních souborů na podporu budoucích ochranných přístupů, jako jsou formátová migrace nebo emulace. Normalizace je založena na identifikaci formátů a jejich významných vlastností a na formátově orientovaných postupech (format policies), které pro

[10] Tvůrci systému Archivematica odůvodňují přímé využití souborového systému jeho robustností a prověřenou dlouhodobou trvanlivostí v porovnání s jinými informačními systémy. Současně je to součást i jejich širší strategie LTP: každá vrstva a komponenta systému LTP není odolná vůči rizikům technologického zastarávání stejně tak jako digitální data samotná. Čím méně těchto technologicky náročných vrstev, tím lépe.

každý vstupní formát specifikují cílový formát, akce, nástroje a postupy při generování archivační a přístupové verze zpracovávaných souborů. Cílové formáty pro normalizaci jsou vybrány na základě takových kritérií, jako jsou aktuální doporučení komunity LTP, veřejná dostupnost specifikace formátu a široký výběr volně dostupných open source nástrojů pro jeho tvorbu a prezentaci, nezatíženost formátu licenčními či patentovými omezeními aj. Administrátor dané instalace Archivematicy může ovšem kdykoliv nastavit vlastní preferované formáty a postupy normalizace.

Součástí systému Archivematica je *formátový registr FPR* (Format Policy Registr), který centrálně spravuje producent systému. Tento registr specifikuje a průběžně aktualizuje formátově orientované postupy doporučené na základě soudobého stavu poznání a osvědčených postupů v oblasti digitálního uchování (administrátor systému má možnost tyto centrálně spravované postupy upravovat a doplňovat v lokální kopii registru). Registr FPR je prostřednictvím rozhraní API dostupný a sdílený nejen všemi organizacemi používajícími systém Archivematica, ale i dalšími zájemci a projekty. Je propojen s registrem PRONOM[11] a plánuje se využití služeb i dalších formátových registrů, například UDFR (Unified Digital Format Registry) nebo Planets Core Registry. Využívá tak poznatků a zkušeností celé komunity LTP.

Instituce může využít registr FPR jako nástroj pro podporu a aktualizaci lokálních postupů v rámci své širší koncepce a strategie digitální archivace. Uživatel má volnost ve stanovení vlastních postupů vycházejících z institucionální strategie LTP či formálních nástrojů pro plánování uchovávání, jako je např. PLATO[12]. Avšak samotné vytváření obecných plánů ochrany a jejich provádění nad archivovanými daty již Archivematica v současné verzi neřeší.

2.5 Zpřístupnění (Access)

Archivematica je navržena tak, aby využívala a podporovala integraci s externími systémy všude tam, kde je možné využít stávajících technologií a není nutné budovat vlastní řešení – příkladem jsou již výše zmíněné technologie pro datová úložiště anebo systémy pro správu a zpřístupnění dat. Zákazníci tak mají možnost využívat dál jejich stávající systémy a systém Archivematica s nimi integrovat „jen“ pro zajištění procesů dlouhodobé archivace.

Již během operace příjmu dat do archivu mohou být generovány přístupové verze digitálních objektů zabalené spolu s dalšími informacemi do balíčku DIP. Tyto jsou následně importovány do externího přístupového systému a jeho prostřednictvím jsou dostupné uživatelům. Archivematica poskytuje nástroje pro základní synchronizaci metadat mezi archivním uložením a externím přístupovým systémem. Jako standardní přístupový systém sdružený s Archivematicou je v současnosti nabízen systém AtoM, vyvinutý tvůrci systému Archivematica pro potřeby komunity archivářů (katalogizační archivační systém). Uživatelé ovšem mohou připojit i jiné přístupové systémy. V rámci různých pilotních projektů bylo ověřeno napojení systémů Archivist's Toolkit, ContentDM, DSpace či Fedora (Islandora)[13].

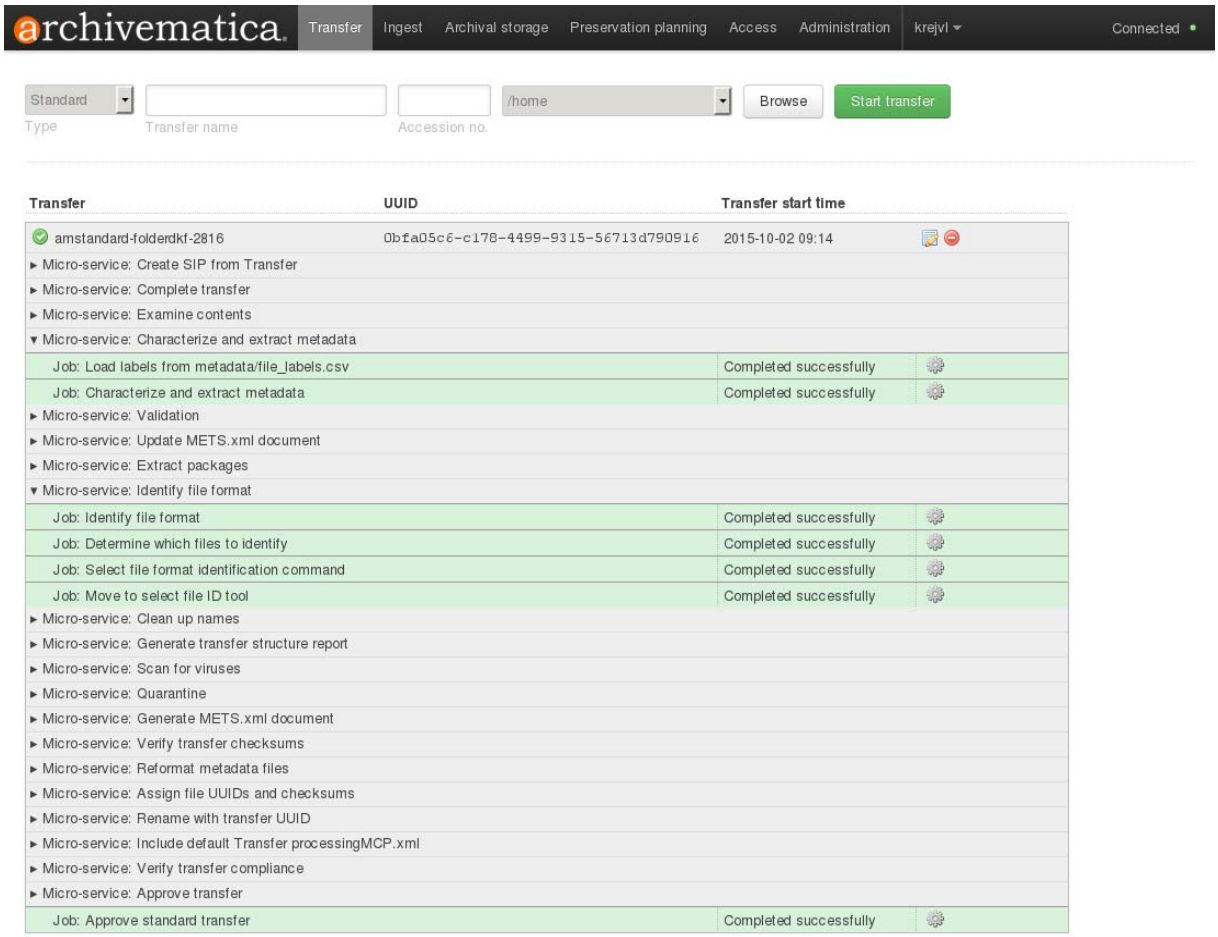
[11] PRONOM byl vytvořen a je provozován Národním archivem Velké Británie. Poskytuje databázi technických informací o souborových formátech a dostupných nástrojích pro ně.
[12] PLATO – The Preservation Planning Tool, http://www.ifs.tuwien.ac.at/dp/plato/intro/
[13] Jedním z příkladů je využití systém Archivematica jako tzv. „dark archive“ pro systém DSpace. Repozitář DSpace slouží uživatelům jako úložný a přístupový systém, k němuž je připojena Archivematica poskytující funkci dlouhodobého archivu. https://www.archivematica.org/wiki/DSpace_integration, https://www.archivematica.org/wiki/DSpace_exports

2.6 Správa archivu (Administration) – Dashboard

Uživatelské rozhraní pro provoz a správu systému/archivu Archivematica se nazývá Dashboard[14]. Je to webová aplikace umožňující zejména

- parametrizovat systém,
- připravovat a přijímat nový obsah do archivu,
- monitorovat a řídit postup zpracování vkládaného obsahu, obvykle formou výběru z předem nastavených variant,
- editovat a doplňovat požadovaná metadata,
- zpracovávat požadavky uživatelů na poskytnutí balíčků AIP,
- poskytovat informace pro plánování uchovávání,
- poskytovat statistické údaje o provozu a operacích systému Archivematica (zatím jen v omezené podobě).

Funkce systému Archivematica jsou rozděleny v souladu s modelem OAIS do služeb popsaných výše, jimž odpovídají jednotlivé části (záložky) nástroje Dashboard – tj. Transfer, Ingest, Archival Storage, Preservation Planning[15], Access, Administration, viz obrázek 3.



Obr. 3 Dashboard – uživatelské rozhraní pro práci s Archivematicou

Při provádění jednotlivých operací zobrazuje Dashboard seznam právě realizovaných mikroslužeb a dle potřeby generuje upozornění tam, kde je nezbytný ruční zásah uživatele –

[14] V původním významu „palubní deska“ nebo „řídicí panel“.
[15] Poskytuje přístup k registru FRP.

např. volba dalšího postupu z více možných variant nebo ošetření chybového stavu. Je ovšem možné nakonfigurovat jednotlivé procesy i tak, aby ruční zásahy byly (přinejmenším z větší části) automatizovány.[16]

2.7 Dlouhodobá správa chráněného obsahu

Současná verze systému Archivematica umožňuje vytvářet kvalitní a robustní archivační balíčky AIP, poskytuje však jen minimum nástrojů potřebných pro jejich dlouhodobou správu. Například v delším časovém úseku může být nezbytné změnit obsah uložených balíčků AIP v souvislosti s migrací zastaralých archivačních formátů na nové formáty, či z důvodu aktualizace metadat uložených v AIP. Alespoň částečné vylepšení v tomto směru by měly přinést další verze systému, kde mezi nově připravovanými funkcionalitami[17] je i vytváření verzí informačních balíčků a operace AIP-reingest. To by mělo umožnit provádět jak drobné aktualizace AIP (například přidání souboru, který chyběl v původním balíčku SIP), tak rozsáhlé velkoplošné změny (například periodické migrace formátů normalizovaných souborů).

Další ze zatím nepodporovaných funkcí jsou replikace balíčků AIP do většího počtu geograficky distribuovaných úložišť a také periodické kontroly integrity těchto balíčků[18]. Uživatelé současné verze systému Archivematica musí využít pro tyto účely externí systémy a nástroje.

3. Další vlastnosti systému Archivematica

3.1 Software Archivematica a mikroslužby

Archivematica je naprogramována v programovacím jazyce Python. Programový kód systému Archivematica, vývojové prostředí a dokumentace jsou volně dostupné a šířené pod licencí AGPL 3.0 (GNU Affero General Public License) a Creative Commons. Systém je připraven k instalaci pod operačním systémem Ubuntu. Alternativně lze připravit jeho distribuci ve formě virtuálního zařízení (appliance) s „přibalenou“ upravenou linuxovou distribucí Xubuntu a sadou open source softwarových nástrojů. Pomocí vhodné virtualizační aplikace (např. Oracle VirtualBox, VMWare Player) může být toto virtuální zařízení se systémem Archivematica provozováno na jakémkoli hardwaru a operačním systému, včetně běžných stolních počítačů. Obraz disku použitý pro virtuální zařízení může být použit také pro vytvoření spustitelného USB disku či DVD nebo pro přímou instalaci systému Archivematica na fyzický hardwaru serverů a pracovních stanic.

Jak bylo uvedeno již dříve, Archivematica využívá koncepci mikroslužeb. To znamená, že informační balíčky vložené do systému Archivematica jsou zpracovávané postupně po dílčích krocích prostřednictvím jednotlivých mikroslužeb zřetězených tak, že výstup jedné je vstupem následující. Každá mikroslužba představuje obvykle několik kroků (jobs) a je implementována jako kombinace skriptů systému Archivematica a jednoho či více volně

[16] Příklady automatizace vybraných postupů lze nalézt například v https://github.com/mjordan/archivematicaworkshop

[17] Viz též https://www.archivematica.org/wiki/Development_roadmap:_Archivematica

[18] Aktivní periodické kontroly integrity jsou plánovány pro některou z budoucích verzí systému Archivematica.

dostupných softwarových nástrojů. Každý z předem instalovaných nástrojů je možné vyměnit relativně snadno za jiný, aniž by to ohrozilo fungování systému jako celku[19].

Při úvodních analýzách a rozpracování funkčního modelu OAIS do uživatelských scénářů identifikovali tvůrci systému původně 24 mikroslužeb sdružených do devíti procesních kategorií[20]:

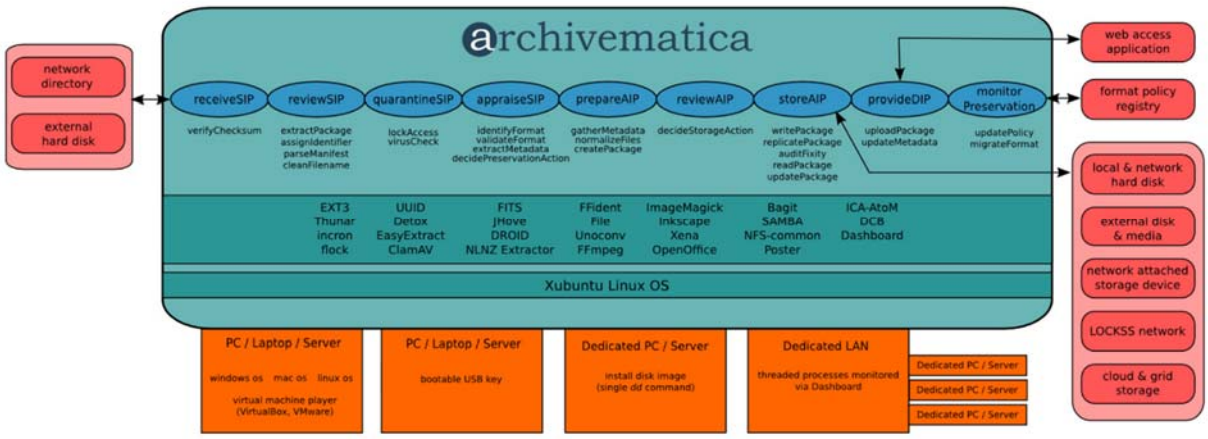
Kategorie	**mikroslužba**
1. receiveSIP	verifyChecksum
2. reviewSIP	extractPackage assignIdentifier parseManifest cleanFilename
3. quarantineSIP	lockAccess virusCheck
4. appraiseSIP	identifyFormat validateFormat extractMetadata decidePreservationAction
5. prepareAIP	gatherMetadata normalizeFiles createPackage
6. reviewAIP	decideStorageAction
7. storeAIP	writePackage replicatePackage auditFixity readPackage updatePackage
8. provideDIP	uploadPackage updateMetadata
9. monitorPreservation	updatePolicy migrateFormat

Rozsah a specifikace mikroslužeb jsou během vývoje systému Archivematica neustále rozšiřovány a upřesňovány, takže aktuální repertoár mikroslužeb je již mnohem širší a propracovanější.

Architektura systému Archivematica využívající mikroslužeb, které jsou implementovány s využitím volně dostupných nástrojů, je znázorněna na obrázku 4.

[19] Například stávající mikroslužba „Scan for viruses“ používá pro antivirovou kontrolu souborů nástroj ClamAv. Pokud bychom ho chtěli vyměnit za jiný antivirový program, stačí adekvátně upravit příslušné skripty připravující vstupy pro kontrolu novým nástrojem.

[20] VAN GARDEREN, Peter. Archivematica: Using micro-services and open-source software to deliver a comprehensive digital curation service. In: iPRES2010: 7th International Conference on Preservation of Digital Objects, September 19 – 24, 2010, Vienna, Austria



Obr. 4 Architektura systému Archivematica[21].

3.2 Standardy

Archivematica využívá řadu zavedených otevřených standardů pro metadatový popis, identifikaci a integraci údajů. Mezi ty nejdůležitější patří:

BagIt[22] – specifikace pro zabalení adresářů se soubory do jednoho balíčku pro dlouhodobé uložení nebo přenos. Důležitou vlastností je generování a zaznamenání kontrolního součtu pro každý jednotlivý soubor v balíčku, což velmi usnadňuje kontrolu integrity a neměnnosti souborů. Archivematica ukládá ve standardu BagIt informační balíčky AIP; současně je schopna přijímat také v tomto formátu balíčky vytvořené jinými systémy.

METS (Metadata Encoding and Transmission Standard)[23] – standard v podobě XML Schema pro zápis všech metadatatových záznamů (popisných, administrativních, strukturálních), a také souborů tvořících digitální objekt, do jednoho XML souboru. Archivematica využívá standard METS pro seskupení všech metadat vytvořených pro množinu souvisejících objektů do jednoho souboru. Tento METS soubor je pak spolu s původními a normalizovanými datovými soubory součástí informačních balíčků AIP.

PREMIS (PREservation Metadata: Implementation Strategies)[24] – standard archivačních metadat, který mimo jiné poskytuje slovník pro zachycení historie (změn) digitálního objektu během jeho uchovávání v archivním systému. Zaznamenává události týkající se digitálního objektu (např. příjem do systému, provedení antivirové kontroly, konverze mezi formáty, ověřování kontrolního součtu), agenty, kteří dané změny provedli (lidé, programy, organizace) a technické charakteristiky objektu samotného (včetně formátu souboru, velikosti, rozlišení). Archivematica generuje metadata ve standardu PREMIS pro uchovávané objekty a přidává je do METS souborů, které tyto objekty popisují.

UUID (Universaly Unique Identifier)[25] – standard umožňující jednoznačně identifikovat informační objekty v distribuovaných systémech bez nutnosti centrální koordinace.

---

[21] Převzato z http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/vanGarderen28.pdf
[22] https://tools.ietf.org/html/draft-kunze-bagit-10
[23] http://www.loc.gov/standards/mets/
[24] http://www.loc.gov/standards/premis/
Pro skvělý úvodní přehled standardu PREMIS viz B. Lavoie and R. Gartner. Preservation Metadata (2nd edition). DPC Technology Watch Report 13-03, May 2013 – dostupný z http://dx.doi.org/10.7207/TWR13-03.
[25] https://tools.ietf.org/html/rfc4122

Identifikátor UUID je 128bitová hodnota (reprezentovaná pomocí 36 alfanumerických znaků) vygenerovaná tak, aby z praktického hlediska zaručovala jednoznačnost všech identifikátorů. Archivematica používá UUID k identifikaci všech objektů, s nimiž pracuje, včetně souborů, procesů či paměťových lokací.

3.3 Škálovatelnost systému

Archivematica využívá architekturu klient/server, která může být využita v různých konfiguracích pro podporu náročných provozních procesů. Pro dosažení lepšího výkonu a zpracování velkých objemů dat je možné distribuovat jednotlivé služby systému Archivematica mezi více uzlů – procesorů, stejně jako lze použít různých scénářů vícenásobné instalace systému samotného. Instituce může provozovat několik systémů běžících současně – a to buď tak, že každý systém zpracovává jinou množinu úloh (například z toho důvodu, aby výpočtově náročná konverze objemných grafických souborů neblokovala provoz systému jako celku), nebo naopak všechny pracují na stejné úloze (např. při souběžném zpracování velkého množství dat).

3.4 Udržitelnost a další vývoj

Archivematica je open source program vyvíjený a šířený bezplatně s podporou společnosti Artefactual Systems. Zpočátku byl vývoj systému financován organizací UNESCO, v současnosti je podporován producentem a z různých dalších zdrojů, například i tím způsobem, že zákazníci objednají a hradí vývoj určitých specifických funkcionalit, které jsou po dokončení dostupné bezplatně všem. Stejně tak jsou komunitou sdíleny komponenty vyvinuté nezávislými vývojáři. Instituce, které požadují technickou podporu při instalaci a nastavení systému, si mohou toto objednat jako volitelnou placenou službu od společnosti Artefactual Systems.

4. Shrnutí – co Archivematica je a co není

Hlavní rysy systému Archivematica lze stručně shrnout do následujících bodů:

- Jde o bezplatný open source systém vyvíjený společností Artefactual Systems Inc. za podpory rozvíjející se zákaznické a uživatelské komunity.
- Systém je společností Artefactual Systems aktivně rozvíjený; několikrát ročně jsou zpřístupňovány nové verze přinášející novou funkcionalitu a opravy chyb.
- Systém nelze považovat za dokončený; některé důležité funkce zatím chybí a například „odladění“ postupů pro hladký průběh hromadného vkládání velkých objemů dat může být náročné.
- Uživatelé mohou ovlivňovat vývoj systému prostřednictvím sponzorování nových funkcionalit (ty jsou pak bezplatně dostupné všem a je garantován jejich přenos do nových verzí systému) a/nebo podáváním námětů na další vývoj (wishlist). Vzhledem k dostupnosti otevřeného kódu může kdokoliv samozřejmě provádět i vlastní vývoj.
- Systém je flexibilní; je postaven na konceptu mikroslužeb, které využívají osvědčené volně dostupné open source nástroje a standardy pro implementaci velké části funkcí spojených s činností archivu a transformací svěřených digitálních dat.

- Předností systému Archivematica je jeho konfigurovatelnost a přizpůsobitelnost, zejména co se týče konfigurace nástrojů napojených na mikroslužby. Uživatelé tak mohou nastavit systém do určité míry podle svých konkrétních požadavků a postupů. Na druhou stranu může tato vlastnost naopak zvyšovat vstupní bariéru pro nové uživatele systému a nováčky v oblasti digitální archivace.
- Velkou část postupů a operací lze automatizovat a omezit tak množství ruční práce při vkládání dat do archivu.
- Základní ochrannou strategií používanou systémem je v současnosti normalizace dat (na základě formátově orientovaných postupů) a generování kvalitně připravených archivačních balíčků, které mohou být uloženy v nezávislých repozitářích.
- Prostřednictvím formátového registru FPR nabízí systém průběžně aktualizované doporučené postupy vycházející z osvědčených postupů v oblasti a komunitě LTP; současně však umožňuje i lokální konfiguraci postupů podle potřeb a požadavků dané instituce.
- Existuje již řada pokročilých projektů nasazení a využití systému Archivematica v různém prostředí a kontextu; je možné sdílet zkušenosti, nástroje a způsobu nasazení systému[26]. Chybí však zatím rozsáhlejší zkušenosti s velkými instalacemi a dlouhodobým provozem.
- Systém neposkytuje (alespoň ve verzi dostupné v době psaní tohoto článku) všechny funkce dle modelu OAIS; zaměřuje se především na funkce související s příjmem dat a jejich přípravou pro archivní uložení a částečně i zpřístupnění; využívá integraci s externími systémy pro zajištění ostatních funkcí (zejména fyzické uložení, plánování a provádění aktivní ochrany uložených dat, zpřístupnění archivovaných dat uživatelům).
- Systém nabízí nízkonákladové řešení dlouhodobého uchování digitálních informací – je tak možné začít danou problematiku řešit již dnes, i když instituce nemá finanční zdroje na nákup drahých komerčních produktů. Není to ale bez práce: úspěšné zavedení vyžaduje netriviální úsilí a znalosti jak pro zvládnutí systému a nastavení procesů pro konkrétní (specifické) potřeby, tak zejména pro integraci systému do širší infrastruktury pro správu a zpřístupnění digitálních dat využívané příslušnou organizací.
- Pro ty, kteří chtějí využít systém Archivematica pro dlouhodobou ochranu svých digitálních dat, ale nemají odborné kapacity na jeho zavedení, provoz a rozvoj, existují placené hostované služby Arkivum[27], ArchivesDirect[28] a další.

5. Závěr

Archivematica je volně dostupný open source systém pro podporu dlouhodobé archivace digitálních dat, který je v současnosti považován mnohými za nejpokročilejší volně dostupný nástroj svého druhu. Na rozdíl od jiných řešení, která se snaží pokrýt všechny funkce týkající se správy, uchování a zpřístupnění dat v rámci jednoho integrovaného systému (např. systém RODA[29]), je Archivematica koncipována jako doplněk stávající infrastruktury. Soustřeďuje se na procesy a služby dlouhodobého uchování a předpokládá integraci s dostupnými externími systémy pro zajištění ostatních funkcí správy dat (správy sbírek, fyzické uložení dat,

[26] Vlastní LTP řešení na bázi systému Archivematica vyvíjí například i Národní archiv ČR.
[27] http://arkivum.com
[28] http://www.archivesdirect.org
[29] http://www.roda-community.org/

zpřístupnění archivovaných dat). Cílem systému Archivematica tak není nahradit používané systémy pro správu a zpřístupnění obsahu, ale doplnit jejich funkcionalitu v oblasti dlouhodobé archivace.

Archivematica je relativně mladý systém vyvíjený společností Artefactual Systems Inc. od roku 2008. Jeho vývoj není dokončen a některé funkce známé z komerčních produktů v systému zatím scházejí. Avšak dynamický rozvoj systému, jeho flexibilita a rostoucí uživatelská komunita z něj vytváří nadějnou alternativu zejména pro ty, kteří hledají perspektivní otevřené řešení a/nebo disponují jen omezenými finančními zdroji. Vedle řady zahraničních projektů a prvních provozních instalací je systém v poslední době aktivně ověřován i v českém prostředí. Je připravováno jeho využití v Národním digitálním archivu vyvíjeném Národním archivem ČR, proběhlo intenzivní testování systému a jeho možností v projektu LTP-pilot podpořeném Fondem rozvoje CESNET, s využitím systému Archivematica počítá také projekt „ARCLib – komplexní řešení pro dlouhodobou archivaci digitálních (knihovních) sbírek“ připravený a podaný několika knihovnami do programu Ministerstva kultury NAKI II na období 2016–2020[30].

Literatura

[1] Archivematica [online]. Artefactual Systems Inc., 2015 [cit. 2015-09-28]. Dostupné z: http://www.archivematica.org/

[2] VAN GARDEREN, Peter a Courtney C. MUMMA. Realizing the Archivematica vision: delivering a comprehensive and free OAIS implementation. In: *iPRES2013: proceedings of the 10th International Conference on Preservation of Digital Objects, 3-5 September 2013, Lisbon, Portugal* [online]. Lisbon: Biblioteca Nacional de Portugal, 2013 [cit. 2015-09-28]. Dostupné z: http://purl.pt/24107/1/iPres2013_PDF/Realizing%20the%20Archivematica%20vision%20delivering%20a%20comprehensive%20and%20free%20OAIS%20implementation.pdf

[3] VAN GARDEREN, Peter. Archivematica: Using micro-services and open-source software to deliver a comprehensive digital curation service. In: *iPRES2010: 7th International Conference on Preservation of Digital Objects, September 19 – 24, 2010, Vienna, Austria* [online]. Vienna, iPress2010, 2010 [cit. 2015-09-28]. Dostupné z: http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/vanGarderen28.pdf

[4] JORDAN, Mark. Introduction to Archivematica : Material for a workshop on Archivematica. In: *GitHub* [online]. Apr 30 2014 [cit. 2015-09-28]. Dostupné z: https://github.com/mjordan/archivematicaworkshop

[5] SCHUMACHER, Jaime et al. *From Theory to Action: “Good Enough” Digital Preservation Solutions for Under-Resourced Cultural Heritage Institutions: A Digital POWRR White Paper for the Institute of Museum and Library Service*[online]. August 2014 [cit. 2015-09-28]. Dostupné z: http://commons.lib.niu.edu/handle/10843/13610

[30] Výsledky veřejné soutěže NAKI II by měly být známy v závěru roku 2015.

[6] LAVOIE, Brian. *The Open Archival Information System (OAIS) Reference Model: Introductory Guide (2nd Edition): DPC Technology Watch Report 14-02 October 2014* [online]. Digital Preservation Coalition, 2014 [cit. 2015-09-28]. Dostupné z: http://dx.doi.org/10.7207/TWR14-02

[7] ČSN ISO 14721. *Systémy pro přenos dat a informací z kosmického prostoru – Otevřený archivační informační systém – Referenční model.* Praha: Úřad pro technickou normalizaci, metrologii a státní zkušebnictví, 2014. 98 s. Třídící znak 31 9620.

[8] MITCHAM, Jenny et al. *Filling the Digital Preservation Gap: A Jisc Research Data Spring project: Phase One report – July 2015* [online]. University of York, University of Hull, 2015. [cit. 2015-09-28]. Dostupné z: http://dx.doi.org/10.6084/m9.figshare.1481170