Uchování digitálního dědictví a systém Archivematica

Článek se zabývá nízkonákladovým přístupem k problematice dlouhodobého uchování digitálních informací a open-source systémem Archivematica. Přináší výsledky uživatelského a systémového hodnocení tohoto systému, které bylo provedeno v rámci projektu LTP-pilot, financovaného z Fondu rozvoje akademického sdružení CESNET. V závěru zmiňuje navazující projekt ARCLib zaměřený na vytvoření uceleného systému a metodiky pro uchování digitálních informací v českých knihovnách a dalších paměťových institucí.

1. Uchování digitálních informací

Stále více institucí začíná znepokojovat otázka, zda jsou schopny zajistit uchování svých cenných digitálních informací pro budoucí pokolení. Budou naše digitální dokumenty a data, obrazové soubory, audiovizuální materiály, databáze či počítačové simulace dostupné a použitelné i za desítky a stovky let? Tato otázka trápí nejen klasické paměťové instituce jako knihovny a archivy, ale dnes již každého, kdo vytváří digitální informace dlouhodobého charakteru. Na rozdíl od tištěných materiálů, které dokáží samy přežít po dlouhou dobu, aniž by se o ně kdokoliv staral, u digitálních objektů toto neplatí. Problémem je nejen krátká životnost a křehkost záznamových médií, ale zejména technologický pokrok. Rychlý vývoj v oblasti informačních technologií vnímáme jako něco pozitivního, co nám přináší stále nové a lepší možnosti. Z hlediska uchování informací znamená ale velkou komplikaci. S příchodem nových technologií postupně mizí starší zařízení, která potřebujeme pro přečtení našich dávno nahraných médií. Zastarává a vytrácí se však i původní software – programy, které jsou nezbytné pro dekódování a zobrazení digitálního záznamu stejně jako pro fungování paměťových zařízení a počítačů. A ke všemu se vytrácí také znalosti potřebné k používání dřívějších technologií a porozumění uchovanému obsahu. Mají-li za této situace digitální informace přežít, musí se o ně někdo systematicky a dlouhodobě starat.

Problematikou dlouhodobého uchování digitálních informací či digitální archivace (Digital Preservation, LTP – Long-Term Preservation) se začaly již před delší dobou zabývat velké knihovny a archivy ve vyspělých zemích, které shromáždily rozsáhlé digitální sbírky a praktické zkušenosti s řešením problémů spojených s technologickým zastaráváním. Mezi lídry patří americká Kongresová knihovna, Britská národní knihovna, Nizozemská národní knihovna či Národní archiv Nového Zélandu. V České republice a na Slovensku se otázka dlouhodobého uchování začala řešit v návaznosti na masové digitalizační projekty v posledních letech: v Národní knihovně ČR vzniklo řešení v podobě LTP úložiště, Univerzitná knižnica v Bratislave zprovoznila Centrálny dátový archív. Všechna tato „velká řešení" se opírají buď o rozsáhlé systémy vyvinuté na míru jednotlivým zadavatelům, nebo o drahá komerční řešení.

Finanční náročnost spolu s velkou šíří nezbytných expertních znalostí a stále ještě nízkým stupněm porozumění problematice vedla po dlouhá léta k přesvědčení, že digitální archivace je výhradní doménou jen těch největších specializovaných organizací. Situace se však mění. Objevila se řada pokročilých nástrojů, technologií a přístupů, které ukazují, že i v malých institucích a s omezenými zdroji lze začít digitální uchování řešit a snižovat tak budoucí rizika. Příkladem tohoto přístupu je projekt POWWR – Preserving digital Objects with Restricted Resources [1]. Jedním z nových nástrojů, které se objevily v posledních letech a vzbudil v odborné komunitě nemalá očekávání, je také systém Archivematica [2].

Archivematica je volně dostupný open-source systém pro digitální archivaci vyvíjený kanadskou společností Artefactual Systems Inc. ve spolupráci s řadou dalších vládních i nevládních institucí po celém světě. Je navržen a implementován s využitím principu mikroslužeb. Ten spočívá v tom, že složité zpracování vstupních dat je rozděleno do série malých vzájemně navazujících úloh, při němž výstup jedné úlohy slouží jako vstup pro úlohu následující. Systém založený na koncepci mikroslužeb je odolnější vůči technologickému zastarávání, neboť vyměnit jednu mikroslužbu za jinou (lepší, novější) je mnohem snazší, než přeprogramovat celou rozsáhlou aplikaci. Většina mikroslužeb systému Archivematica je navíc implementována pomocí existujících volně dostupných nástrojů (jako jsou nástroje pro identifikaci a kontrolu souborových formátů, extrakci metadat, migraci zastaralých formátů aj.), takže systém se opírá o rozsáhlé a rychle se rozvíjející know-how celosvětové LTP-komunity.

2. Projekt LTP-pilot

Systém Archivematica vzbudil pozornost i u řady akademických organizací v České republice. Ústav výpočetní techniky Masarykovy univerzity připravil proto spolu s Moravskou zemskou knihovnou v Brně projekt, jehož cílem bylo důkladněji prověřit možnosti tohoto systému. Projekt s názvem „Pilotní projekt pro low-barrier přístup k ochraně digitálního obsahu (LTP-pilot)" byl realizován s podporou Fondu rozvoje akademického sdružení CESNET v roce 2015.

Hlavním záměrem projektu bylo ověření funkcí, možností a případných omezení systému Archivematica pro podporu dlouhodobé archivace digitálních informací jako nízkonákladové alternativy k obvykle drahým komerčním řešením a získání zkušeností pro další rozvoj problematiky digitálního uchování v České republice. Realizace projektu se zaměřila na pět hlavních oblastí: 1. systém a jeho vývoj (architektura a systémové vlastnosti systému Archivematica); 2. úložiště (napojení systému Archivematica na různé typy datových úložišť a zajištění bit-level ochrany a integrity archivovaných dat); 3. uživatelské scénáře (návrh postupů pro testování systému Archivematica s ohledem na standardy a požadavky provozu dlouhodobého digitálního archivu a simulování praktických událostí při provozu archivu); 4. digitální kurátorství (uživatelské testování systému pro různé typy digitálních dat s využitím připravených scénářů); 5. standardy (zmapování standardů a doporučené praxe v oblasti digitálního uchování a posouzení funkcionality a datového modelu systému Archivematica z hlediska standardů ISO 14721 – OAIS a ISO 16363 – Audit a certifikace).

Odborná veřejnost byla seznámena s výsledky projektu na workshopu, který se uskutečnil 1. prosince 2015 v Národním archivu v Praze. Všechny písemné výstupy projektu (celkem 18 dokumentů) jsou veřejně dostupné na webu *http://ltp-po-*

rtal.cz/ [3]. V další části tohoto příspěvku se proto zaměříme jen na přehledová shrnutí závěrů k systému Archivematica, která byla součástí závěrečné zprávy projektu pro Fond rozvoje CESNET. Předpokládáme, že čtenář je alespoň rámcově obeznámen s referenčním modelem dlouhodobého archivu OAIS (Open Archival Information System) [4], na který je v dalším textu odkazováno.

3. Systém Archivematica a jeho zhodnocení

Systém Archivematica, zmíněný již výše, je relativně mladý produkt, který vešel do širšího mezinárodního povědomí teprve před několika lety (beta verze uvolněna v roce 2009) a vyvolal značný zájem a velká očekávání. Obecný popis systému a jeho přehlednou charakteristiku lze nalézt například v [5]. V porovnání s některými jinými obdobnými systémy, které vznikly spíše jako prototypy výzkumných projektů a dlouhodobě se nerozvíjely, udržela Archivematica dynamiku svého rozvoje a získává postupně nové uživatele. Doposud dostupné informace o systému jsou však poměrné kusé a chybělo i jeho hlubší a nezávislé posouzení. Snaha napomoci alespoň částečně tyto mezery zacelit a poskytnout širší české a slovenské komunitě kvalifikovanější obrázek byla jednou z hlavních motivací pro vznik projektu LTP-pilot.

Systém Archivematica lze všeobecně hodnotit jako perspektivní rozvíjející se systém, který lze nasadit již dnes pro některé dílčí činnosti digitálního uchování, který ale nepokrývá všechny funkce dle modelu OAIS a ne vždy dosahuje zatím potřebného stupně zralosti. Na druhou stranu je rozvoj systému poměrně dynamický (každoročně je uvolňováno i několik nových verzí)[1], takže lze přepokládat vývoj směrem k vyšší kvalitě i širší funkcionalitě. Tvůrci systému od počátku deklarují, že Archivematica není koncipována jako ucelené řešení „pro vše", ale jako archivační doplněk systémů pro správu a zpřístupnění dat využívaných danou institucí.

Archivematica – uživatelský pohled

Z pohledu uživatelů (správců digitálního obsahu, digitálních kurátorů) je Archivematica ve skutečnosti především systém použitelný pro první zpracování archivních dat, tedy pro vložení digitálních dat (operace Transfer), jejich konverzi na vstupní informační balíček SIP (operace Ingest), aplikaci formátových politik a následně konverzi těchto dat do archivního informačního balíčku AIP. Archivní informační balíček je nezávislý na technologii repozitáře a používá otevřené metadatové standardy (METS a PREMIS) a balíčkovací standard BagIt. Archivní balíček obsahuje všechny typy informací vyžadované informačním modelem OAIS; kromě popisných informací a vlastní obsahové informace zahrnuje také ochranná metadata (fixity, provenance, context, reference information) a informace o formě reprezentace.

Archivematica bývá někdy prezentována jako kompletní systém pro digitální uchování – ve skutečnosti jím ale (přinejmenším zatím) není. Z pohledu funkcionality OAIS je systém Archivematica, tedy její jádro, v podstatě propracovaný modul Ingest. Implementace funkcí potřebných pro zpřístupnění archivovaných dat koncovým uživatelům a funkce správy dat v komplexních situacích životního cyklu dat musí byt realizovány vně systému, stejně jako plánování dlouhodobého uchovávání probíhající s nějakým časovým odstupem po vložení dat.

Systém sice poskytuje zdání, že má i moduly Administrace, Plánování ochrany, Archivní uložení aj., jak by odpovídalo OAIS, ale skutečnost je jiná. Např. Archivní uložení je pouze prostředí, kde lze procházet uložené balíky AIP. Možnosti vyhledávání jsou velmi omezené. Modul Administrace zastřešuje nastavení různého typu, od minimalistické správy uživatelů (role pouze uživatel nebo administrátor), nastavení ukládacích lokací až např. po přehled problémů v průběhu zpracování. Důležitou součástí Administrace je nastavení jednotlivých kroků pro operace Transfer a Ingest, které jsou pak používány jako implicitní pro všechna vstupní data. Nejpropracovanější částí systému Archivematica a vlastně jeho srdcem je (vedle funkcí Transfer a Ingest) modul Plánování ochrany, i když jeho název je poněkud zavádějící. V porovnání se zavedenými LTP-systémy chybí mnohé funkce, mimo jiné vytváření různých scénářů akcí dlouhodobé ochrany, specifikace vlastností, které je potřeba uchovat, či možnosti porovnání výsledků testovacích scénářů. Plánování ochrany v pojetí systému Archivematica je v podstatě formátová knihovna FPR – Format Policy Register, která obsahuje seznam formátů, nástroje na různé aktivity a pravidla pro zpracování souborů. Důležitá jsou zejména pravidla a nástroje na tzv. normalizaci, tedy migraci konkrétních vstupních formátů dat, které nejsou považovány za vhodné pro dlouhodobou ochranu/uchování, do jiných vhodnějších formátů.

Systém nenabízí správu archivních dat. Vyhledávání v rámci systému Archivematica je minimalistické a použitelné jen jako nouzové. Editace obsahu nebo metadat již hotových balíčků AIP není zatím možná. Pokud je toto potřeba, je nutno kompletně balíček AIP vložit znovu s potřebnou změnou. Pro správu a zpřístupnění dat, tedy AIP balíčků a jejich metadat, je nutno připojit některý externí systém pro správu digitálních objektů. Tvůrci systému Archivematica nabízí implicitně pro tyto účely vlastní open-source produkt AtoM (vyvinutý pro potřeby archivářské komunity), lze ale využít propojení i na některé široce používané repozitářové systémy jako např. DSpace, ContentDM či Fedora/Islandora.

Obecné nastavení systému Archivematica není obtížné. Nastavit po instalaci systém pro první transfer a ingest není z pohledu správce systému otázka dní, ale spíše hodin. Provést první zpracování dat lze pro jednoduché (ne komplexní a validní) soubory v podstatě ihned – a to díky implicitním přednastaveným pravidlům a nástrojům. Před ostrým spuštěním systému je však nutno detailně analyzovat a případně upravit formátovou knihovnu FPR. Je nutné zjistit, jaké nástroje jsou připojeny pravidly ke konkrétním formátům a zda jejich výstupy odpovídají potřebám instituce.

Testování ingestu ukázalo, že Archivematica je schopná přijímat jak jednotlivé soubory, tak celé sady dat z určené lokace na serveru. Transfer/ingest jednotliviny nebo většího množství dat ve struktuře BagIt se také ukázalo jako bezproblémové. Nedostatkem BagIt je však nemožnost přidat metadata k jednotlivým souborům v obsaženým balíku. Problémy se ukázaly při příjmu poškozených, nevalidních a jinak problematických souborů. Systém často chyby neidentifikoval, pokud je identifikoval, transfer/ingest přesto dále probíhal a byl i vytvořen archivní balíček, což odporuje logice obvyklé u jiných LTP-systémů.

Výše uvedené nedostatky a omezení systému Archivematica jsou, jak mimo jiné ukazují příklady některých zahraničních projektů, technicky řešitelné. Chybějící nebo nedostatečně rozvinutou funkcionalitu v různých oblastech – vyplývajících jak z požadavků odvozených z OAIS, tak z praxe paměťových institucí – lze doplnit integrací s jinými systémy, úpra-

[1] Vývoj systému – posuzováno frekvencí vydávání nových verzi – se v poslední době zpomalil. Zatímco dříve vycházelo i několik nových verzí ročně, aktuálně nejnovější a dlouho avizovaná verze 1.5 se připravuje prakticky již celý rok, od května 2015

vami mikroslužeb, úpravami datového modelu AIP nebo zadáním požadavku na vývoj nových funkcí. Některé ze zmíněných nedostatků, například chybějící funkce pro aktualizaci metadat AIP nebo aktualizace balíčku AIP jako celku, jsou plánované a sponzorované funkce ve vývojovém plánu systému. V dalších oblastech, například zpřístupnění a vyhledávání, se lze při nasazení inspirovat projekty, kde byla Archivematica integrována s jinými systémy a/nebo hledat vlastní lokální řešení. Rozsáhlejší projekty národního rozsahu, předpokládající nasazení systému v komplexním prostředí s mnoha dodavateli dat, vyžadují pečlivou revizi kódu jednotlivých mikroslužeb. Přitom, ale i takový zásah využívá existující architektury a filosofie systému. S dalšími úpravami je potřeba počítat při plánování nasazení formou služby dalším institucím. Čili zjištěná omezení a funkční nedostatky jsou vždy nějakým způsobem řešitelné. Zájemce si jich však musí být vědom a v projektu nasazení systému Archivematica v produkčním prostředí musí počítat s nutností integrace dalších systémů, stejně jako s určitým rozsahem nového vývoje nebo přizpůsobení.

Archivematica – systémový pohled

Z hlediska koncepčního návrhu a realizace je možné hodnotit systém Archivematica pozitivně. Realizace jádra a napojení na mikroslužby využívající dostupné externí LTP-nástroje je řešeno, s ohledem na cíle, kterých chce systém dosáhnout, rozumně a vhodně i pro budoucí vývoj. Ze systémového pohledu se jeví vhodné i oddělení jádra systému, služby Storage Sevice (pro fyzické ukládání archivovaných dat) a přístupového systému (např. AtoM) do samostatných modulů, které je možno provozovat na oddělených serverech. Mikroslužby lze vytvářet v jakémkoli programovacím jazyce, jehož výstup je spustitelný na serveru systému Archivematica.

Požadavky na výkon hardwaru definované v oficiální dokumentaci se nevymykají parametrům běžného menšího serveru. Při reálném provozu závisí potřeba systémového výkonu velmi na konkrétním nastavení procesu zpracování (jaké mikroslužby a které externí nástroje jsou použity) a jaká data se na vstupu zpracovávají. Archivematica používá pro řízení a spouštění mikroslužeb systém Gearman, který se stará o vysokou paralelizaci spouštěných mikroslužeb. Ve výsledku se tak při zpracování využívá v maximální možné míře všech jader procesoru. Vkládat vstupní data a zahajovat proces transferu je možné nezávisle na aktuálně běžících procesech zpracování dříve vložených dat. Zpracování dat je tak možné paralelizovat i na úrovni správy vstupu dat spuštěním více transferů najednou.

Prvotní instalace systému Archivematica podle dokumentace je bezproblémová a dá se zvládnout v řádu jednotek hodin (počítáno i s instalací a nastavením samotného serveru s linuxovou distribucí Ubuntu). Vyplatí se instalovat systém na doporučené verze operačního systému; jiné varianty (byť i avizované jako možné) nejsou vhodné – z hlediska provozu se ukazují jako problémové. Určité problémy jsme zaznamenali při upgradech již zavedeného systému na novější verze.

Největší slabinou z hlediska systémové správy a lokálního vývoje se ukazuje být chybějící či neúplná a nepřesná dokumentace. Dokumentace na oficiálním webu je sice průběžně doplňována, ale pro expertní zvládnutí systému je nedostatečná. Příkladem může být rozhraní REST API, které není dosud dokumentováno vůbec, nicméně pro automatizaci vkládání dat do archivačního systému jde o zcela klíčovou komponentu. Dalším příkladem, kde nedostatečná dokumentace velmi znesnadňuje lokální rozvoj a přizpůsobování systému, je vývoj a nastavování vlastních mikroslužeb. Ačkoliv koncept mikroslužeb a využití externích nástrojů nabízí možnost velké flexibility systému, v praxi je při nedostatečné dokumentaci vývoj vlastních mikroslužeb proces časově náročný.

V průběhu testování se systém ukázal jako málo robustní (projekt testoval verzi systému 1.4, aktuálně před zveřejněním je již nová verze 1.5). Narazili jsme také na některé limity nebo chyby systému (např. limity na počet objektů ve vkládaném balíčku, způsob práce s nevalidními soubory apod.) Tyto problémy ukazují na to, že míra rozšíření a uživatelská komunita nejsou zatím tak velké a že systém je třeba pro ostré použití dobře odladit.

4. Jak dále v nízkonákladové digitální archivaci

Projekt LTP-pilot si nekladl za cíl vytvořit a nasadit funkční řešení pro digitální archivaci. Připravil však podmínky pro navazující rozsáhlejší projekt, který již tyto ambice má. Tím je projekt „ARCLib – komplexní řešení pro dlouhodobou archivaci digitálních (knihovních) sbírek" přijatý v závěru roku 2015 k řešení v rámci programu Ministerstva kultury ČR NAKI II (Program na podporu aplikovaného výzkumu a experimentálního vývoje národní a kulturní identity). Jde o 5letý projekt na období 2017 –2020, jehož řešitelskými organizacemi jsou Knihovna Akademie věd ČR, Masarykova univerzita, Moravská zemská knihovna v Brně a Národní knihovna ČR. Smyslem projektu ARCLIb je navázat na dosud získané poznatky a vytvořit komplexní LTP řešení na bázi open-source a metodiky pro potřeby dlouhodobé ochrany digitálních dat pro knihovny v ČR, s možností využití i v dalších paměťových institucích (archívy, muzea, galerie). Řešení projektu bylo zahájeno v březnu 2016; vedle výzkumu a dalšího prohloubení znalostí z problematiky digitálního uchovávání má projekt definovány čtyři hlavní cíle a praktické výstupy: 1. vývoj komplexního LTP open-source řešení ARCLib pro digitální archivaci s využitím systému Archivematica a dalších vhodných technologií; 2. vytvoření metodiky pro dlouhodobou logickou ochranu digitálních dat pro české prostředí s ohledem na mezinárodní standardy (zejména referenční model OAIS – ČSN ISO 14721 a ČSN ISO 16363); 3. vytvoření metodiky a návrh řešení pro fyzické ukládání velkého množství dat a zajištění bit-level ochrany pro potřeby dlouhodobé archivace; 4. ověření vytvořených řešení v praxi formou poloprovozu.

Poděkování

Závěry k systému Archivematica publikované v tomto článku vychází z rozsáhlého testování a studií, na kterých se podílel širší řešitelský tým projektu LTP-pilot (FR CESNET, č. 516R1/2014). Celému týmu patří proto mé poděkování. Jmenovitě jde o následující kolegy: Vlastimil Krejčíř, Michal Růžička, Lukáš Hejtmánek a Michal Konečný (všichni Masarykova univerzita), Daniel Sečík, Zdeněk Hruška, Petr Žabička a Pavla Rychtářová (Moravská zemská knihovna v Brně), Marek Melichar a Andrea Miranda (Univerzita Karlova v Praze), Jan Hutař (Archives New Zealand), Rudolf Vohnout (Jihočeská univerzita v Českých Budějovicích).

Použitá literatúra

[1] From Theory to Action: "Good Enough" Digital Preservation Solutions for Under-Resourced Cultural Heritage Institutions [online]. A Digital POWRR White Paper for the Institute of Museum and Library Services, August 2014. Northern Illinois University. [cit. 2016-05-02]. Dostupné z *http://commons.lib.niu.edu/handle/10843/13610*

[2] Archivematica [online]. Artefactual Systems Inc., 2015 [cit. 2016-05-02]. Dostupné z: *http://www.archivematica.org/*

[3] 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.

[4] LTP Portál.cz – výstupy projektu LTP-pilot [cit. 2016-05-02]. Dostupné z *http://ltp-portal.cz/*

[5] BARTOŠEK, Miroslav. Archivematica – open source systém pro digitální archivaci. Knihovna: knihovnická revue, 2015, 26(2), s. 25-38. ISSN 1801-3252. Dostupné také z: *http://knihovnarevue.nkp.cz/*

RNDr. Miroslav Bartošek, CSc.

bartosek@ics.muni.cz

(Masarykova univerzita, Brno; Ústav výpočtovej techniky MU)