Archivematica - svobodný systém na ochranu digitálních dokumentů

Robert Šiška (5. sem., 3. roč.)

5. prosince 2014

http://www.archivematica.org/,
Artefactual Systems Inc.

1 Co je Archivematica

Archivematica je svobodný software vystavěný podle referenčního modelu a ISO standardu OAIS (Open Archival Information System). Zabývá se řešením některých aspektů ochrany digitálních dokumentů s důrazem na urdržování srozumitelnosti dokumentů automatickými konverzemi formátů. Není to monolitický software, ale sada mnoha dílčích programů, které umožňují postupné zpracování základních entit OAIS modelu pomocí dílčích operací - přístupem, který autoři projektu nazývají mikro-služby. Projekt je distribuován spolu s webovou aplikací ICA-AtoM od stejných autorů umožňující webový přístup k archivovaným objektům, může být však kombinován s ostatními systémy určenými k přístupu k dokumentům.

2 Cíle projektu

Projekt vyvíjí a spravuje kanadská společnost Artefactual Systems ve spolupráci s technickým výborem UNESCO programu „Paměť světa“ (anglicky

Memory of the World), Museem moderního umění v New Yorku, Rockefellerovým archivačním centrem a mnoha dalšími organizacemi zabývajícími se digitální archivací. Prvotní verze projektu poprvé spatřily světlo světa v roce 2009 a projekt zůstal ve stádiu beta až do roku 2014, kdy vyšla stabilní verze 1.0. Je to tedy velmi mladý projekt, na kterém stále probíhá poměrně aktivní vývoj. Aktuální vývojová verze je 1.3 a její hlavní přínos je integrace s dalším svobodným projektem DuraCloud (software na archivování pomocí cloud technologií).

Hlavním cílem projektu je vytvoření systému pro údržbu digitálního repozitáře, který je přístupný lidem bez technického zázemí, ale který by zároveň splňoval veškeré nároky kladené archivačními problémy, jako konverze formátů, automatické opatřování metadaty, workflow apod. Většina obdobných profesionálních systémů (jako Roda nebo DAITSS) vyžadují nemalé zkušenosti s administrací serverů, které archiváři často nemají.

Dalším cílem je ponechat veškeré svobody uživatelům. Celý systém je k disposici zdarma a zdrojové kódy jsou veřejně dostupné. Vývoj je veden agilními metodami a autoři udržují kontakt s komunitou. Všechny zdrojové kódy systému jsou licencovány pod AGPL a externí software je kontrolován pro licenční kompatibilitu před tím, než je do projektu integrován.

3 Popis projektu a jeho výsledků

3.1 Technické řešení

Archivematica není multiplatformní aplikace. Je pevně svázána s prostředím GNU/Linux, i když by teoreticky měla být funkční na jakémkoliv systému splňující standardní hierarchii souborobého systému (FHS). Předkompilované balíky se nacházejí v repozitářích systému Ubuntu. Další možností je vytvoření živé distribuce na přenosném médiu. Systém je ale také distribuován jako předkonfigurovaný virtuální obraz specializovaného operačního systému (založeného na Ubuntu) a je tak jednoduše použitelný na všech systémech schopných virtualizace. To redukuje celou instalaci na pouhé stažení sou-

boru a jeho spuštění ve virtualizačním nástroji, což z něj činí nejjednodušeji použitelný systém tohoto typu.

Zvolení vícevrstvé a modulární architektury, v nichž jednotlivé části komunikují pomocí síťových protokolů, může být systém nasazen na více strojů, což v kombinaci s vyvažováním zátěže výrazně zvětšuje škálovatelnost.

Software je z převážné napsán v jazyce Python a jednotlivé mikro-služby mohou být napsány v jakémkoliv jazyce, který je umožňuje přístup k souborovému systému a relačním databázím. Mnohé z nich využívají externí programy na normalizaci dokumentů, identifikaci typu souborů, konverzi obrázků, grafiky a videí, atd.

3.2 Architektura

Po vzoru OAIS definuje Archivematica tři základní typy entit. Producentem nahraný dokument je jako SIP (Submission Information Package) vložen do archivačního systému, který jej zpracuje a vytvoří jednu či více variací určené k archivaci - AIP (Archival). Mimo to jsou vytvářeny variace DIP (Dissemination), které jsou určené k veřejnému přístupu.[1] Každý balík obsahuje dokument samotný a dále také kontrolní součty, metadata, záznamy o vložení a další informace. Tyto dodatečné informace jsou pro každý typ entity jiné.

Tento proces je základem modelu OAIS - tzv. *ingest to access*. Archivematica jej implementuje pomocí mikro-služeb. Mikro-služby jsou dílčí operace na entitách, které jsou určené k řetězení. Každá mikro-služba provádí jeden úkol na dané entitě a po skončení je entita zpracována další službou v pořadí.

V jádru všeho je software typu klient-server určený k distrubovanému provádění úkolů. Konfigurace serveru definuje pořadí mikroslužeb pro jednotlivé úkoly (např. vložení SIP, vytvoření AIP/DIP). Mikro-služby samotné jsou součást klientů, kteří informují server o tom, které služby nabízí a čekají, až je server zaměstná. V praxi to znamená, že server sleduje řadu složek, které reprezentují stav procesu. Mikro-služby pak upravují a přesouvají dokument, dokud řetězec neskončí. Nepříjemný důsledek tohoto řešení je, že server i všichni klienti musí mít přístup do stejné složky, což může vytvářet

úzké hrdlo při nadměrné diskové nebo síťové aktivitě. Tento problém lze částečně řešit použitím distribuovaných souborových systémů (např. RAID). O správu a přístup k uložištím se stará samostatná aplikace s vlastním webovým rozhraním. V současnosti však podporuje pouze ukládání do souborového systému; obecné rozhraní úložného systému je zatím ve vývoji. Konfigurace serveru i mikro-služeb je uložena v relační databázi SQL.

Pro ovládání systému slouží uživatelsky přívětivá webová aplikace komunikující se serverem. Je to víceuživatelský systém, který umožňuje konfigurovat a spouštět úlohy, sledovat a ovlivňovat jejich průběh (některé akce můžou čekat na volbu), upravovat metadata, vytvářet statistiky a podobně.

3.3 Politika ochrany dokumentů

Ochrana digitálních dokumentů souhrnně označuje aktivity vedoucí k zajištění použitelnosti digitálních objektů po mnoho let.[2] Pro tento účel umožňuje Archivematica definovat skupiny digitálních formátů a definovat výstupní formáty pro archivaci a pro přístup. Například všechny audio soubory jsou archivovány v bezztrátovém formátu a veřejně přístupny ve formátu MP3. Autoři kladou důraz na to, aby všechny archivační formáty byly svobodné implementace standardizovaných formátů.

Jak se vyvíjejí nástroje a velikost uložišť, mění se i formáty. Politika převodu formátů tedy není nic vytesaného do kamene. K tomuto účelu spravuje společnost Artefactual veřejně dostupný server, který slouží jako strukturovaný seznam pokynů k normalizaci jednotlivých typů dokumentů. Instance systému Archivematica si tedy může ke svému lokálnímu nastavení navíc přidat i nejnovější politiku z centrálního serveru. Plánem vývojářů je kompatibilita s registry PRONOM a/nebo UDFR.

Identifikace typu souboru může být založena na analýze dat pomocí nástroje FITS, nebo pouze podle přípony. Další metody je snadné dodat. Po zjištění typu souboru jsou dokumenty charakterizovány - vytváří se tzv. významné charakteristiky, které jsou potřebné pro dlouhodobé uchování digitálních dokumentů.[3, s.15] To může být třeba barevná hloubka a rozlišení u obrázků,

počet kanálů a frekvence vzorkování u audio souborů a podobně. Tyto charakteristiky jsou vloženy do metasouboru formátu METS archivovaného AIP. Do řetězce mikro-služeb je možné zapojit také třeba přepis pomocí OCR.

4 Závěr

Archivematica ke své činnosti orchestruje celý operační systém a spoustu externích nástrojů a plně tak využívá svého statutu svobodného software. Tímto aspektem se projekt drží dvou pravidel unixové filosofie „dělej jednu věc, ale pořádně“ a „piš programy tak, aby spolupracovaly.“ Tím, že distribuují vlastní operační systém jako virtuální obraz, mají celý systém pod kontrolou a instalaci systému zvládne archivář i bez potřeby větších technických znalostí a to včetně instalace distribuované na více strojů. Jako pozitivní vidím také snahy integrovat existující systémy jako DSpace, ContentDM, Archivist's Toolkit a jiné. Z plánů pro budoucí verze můžeme vidět, že většina práce na projektu je sponzorována univerzitami, knihovnami a dalšími institucemi z celého světa. O Archivematicu se před nedávnem probudil zájem taky v České republice například ze strany MZK. [4]

5 DC Metadata

TITLE=Archivematica - svobodný systém na ochranu digitálních dokumentů
CREATOR=Robert Šiška
DESCRIPTION=Recenze softwaru Archivematica
SUBJECT.Keywords=Archivematica, OAIS
DATE.Created=2014-12-05
LANGUAGE=czech
FORMAT.Medium=application/pdf

Reference

[1] *ISO 14721:2012 Open archival information system (OAIS) – Reference model*. International Organization for Standardization, Geneva, Switzerland.

[2] GLADNEY, Henry M. *Preserving Digital Information*. New York: Springer, 2007.

[3] WILSON, Andrew *Significant Properties of Digital Objects* [online] National Archives of Australia, 2008. Dostupné z: `http://www.dpconline.org/docs/events/080407sigpropsWilson.pdf`

[4] Archivematica v ČR. In: *Digital Preservation CZ - Blog* [online]. Dostupné z: `http://www.digitalpreservation.cz/2014/04/archivematica-v-cr.html`