Digitální archiv Masarykovy univerzity

Popis způsobu uložení archiválií v digitální podobě

verze 1, 2. 6. 2015

1. Úvod

Tento dokument vychází z dokumentu Koncepce dlouhodobého uchovávání a ochrany dokumentů určených pro předávání do Digitálního archivu Masarykovy univerzity (MU). Jeho cílem je upřesnit konkrétní technická řešení navržená a platná ke dni vydání tohoto dokumentu. Na rozdíl od koncepčních návrhů, které mají trvalejší charakter, zřizovatel Digitálního archivu MU předpokládá, že technické zajištění způsobu uložení archiválií může snadněji reagovat na překotné změny, které přináší vývoj v oblasti digitálních technologií.

2. Vnější rozhraní Digitálního archivu MU

Základním komunikačním protokolem pro komunikaci s Digitálním archivem MU ze strany původců archiválií, pracovníků archivu, jeho technických správců a v neposlední řadě badatelů zadávajících dotazy na přístup k digitálním archiváliím je protokol HTTPS nad veřejně přístupnou sítí Internet prostřednictvím rodiny protokolů TCP a IP verze 4. Zabezpečení přístupového kanálu aplikačního protokolu HTTP je prováděno prostřednictvím asymetrické kryptografie protokolem SSL, přičemž certifikát serveru je vystaven certifikační autoritou, obecně přijímanou v nejrozšířenějších internetových prohlížečích (např. TERENA SSL CA).

3. Formát vstupních dat

Digitální archiv MU přijímá data standardně ve formátu XML souboru standardu SIP, který je zároveň validní vůči Národnímu standardu pro systémy elektronické spisové služby (NSESSS).

Druhým způsobem pro příjem dat je jejich předání prostřednictvím webového formuláře ručním vyplněním všech metadat do příslušných formulářových polí.

Formát souborů s digitálními dokumenty je dán jednotlivými původci. Pro účely dlouhodobé archivace jsou však vybrané formáty automaticky konvertovány do formátů pro dlouhodobou archivaci vhodnějších. Konverze však vždy znamená přidání nového souboru v novém formátu k původnímu dokumentu, který Digitální archiv MU také uchovává, aby nedošlo prováděnou konverzí ke ztrátě původní informace. Problematika dlouhodobé archivace jednotlivých formátů je obecně složitý problém, který nelze v době vzniku Digitálního archivu MU plně uspokojivě vyřešit. Proto je Digitální archiv MU připraven další konverze provádět i v budoucnu s tím, že s určitou mírou ztráty informace jsou nyní konvertovány ty formáty, které jsou pro původce z Masarykovy univerzity nejčastější.

Tabulka automatických konverzí Digitálního archivu MU:

Typ souboru	Vstupní formáty	Výstupní formáty
Prostý text	TXT	PDF/A-1
Portable Document Format	PDF	PDF/A-1, TXT
Open Office	ODT, ODS, ODP	PDF/A-1, TXT
Microsoft Office	DOC, DOCX, XLS, XLSX, PPT, PPTX	PDF/A-1, TXT
Google Documents	GDOC	PDF/A-1, TXT
Rastrové typy	JPG, PNG, GIF, PDF	PDF/A-1, TXT (vč. OCR)

Ostatní formáty nejsou konvertovány v rámci Digitálního archivu MU automaticky. Z pohledu původců se jedná o velmi okrajové typy dat, jejichž konverzi je možné zajistit na straně původců. Tyto soubory jsou uchovávány výhradně v původním formátu, s výjimkou souborů, u kterých Digitální archiv MU detekoval přítomnost závadného kódu (viru), který by mohl narušit bezpečnost Digitálního archivu MU nebo jeho uživatelů. Takový soubor je uchován v trezoru s přístupem omezeným pouze pro správce archivu.

4. Formát uložení dat v Aktivní zóně

Data jsou v Aktivní zóně Digitálního archivu MU uložena v samostatných souborech v tom formátu, ve kterém byly do archivu přijaty od původce nebo ve formátu, do kterého byly po vstupu automaticky zkonvertovány. Každý soubor vstupního dokumentu tak může být reprezentován jedním nebo více souborů uložených v archivu. Soubory jsou uloženy ve standardním souborovém systému, jehož formát má otevřený zdrojový kód pro zajištění dlouhodobého přístupu k souborům při použití jiného operačního systému. Souborový systém zvolený pro tento typ dat, který výše uvedené vlastnosti splňuje, je XFS.

5. Formát uložení metadat v Aktivní zóně

Metadata jsou v Aktivní zóně uložena formou strukturované relační databáze. Pro účely zajištění zpětné kompatibility, dostatečného výkonu a odolnosti vůči ztrátě dat a výpadku, byla zvolena proprietární relační databáze Oracle Database. Nevýhodou tohoto řešení na rozdíl od všech ostatních součástí archivu je závislost formátu uložených metadat na konkrétním proprietárním formátu. Vzhledem k tomu, že metadata v aktivní zóně slouží výlučně k aktivnímu přístupu (nikoliv k archivaci), jsou archivována v Archivní zóně a v rámci provozu permanentně v operační paměti jednotlivých procesů relačního databázového serveru, není proprietární formát daného formátu kritický. V případě přechodu na novější verzi, který je prováděn v pravidelných intervalech, provede Digitální archiv MU nejprve export dat z formátu původní verze do přenosného formátu a poté import do nové verze formátu uložených dat. Tento převod je proveden nejpozději ve lhůtě stanovené výrobcem databázového serveru pro zajištění podpory dat v původním formátu. Nadto jsou všechna metadata archivována v rámci Archivní zóny způsobem zajišťujícím otevřený přístup k metadatům i do budoucna.

6. Formát uložení dat a metadat v Archivní zóně

Archivní zóna ukládá dohromady data i metadata o dokumentech resp. spisech archivovaných v rámci Digitálního archivu v samostatných textových souborech typu XML v archivním formátu AIP. Tento způsob zajišťuje otevřený přístup k datům nezávislý na konkrétním výrobci software pro daný formát.

V rámci Archivního diskového subsystému jsou soubory v archivním formátu uloženy pomocí otevřeného souborového systému XFS.

V rámci Archivního páskového subsystému jsou soubory v archivním formátu uloženy pomocí otevřeného systému Linear Tape Open (LTO).

7. Technologické zázemí a aplikace Digitálního archivu MU

Digitální archiv MU je založen primárně na otevřených technologií, tzn. technologiích, ke kterým je veřejně zdarma dostupný zdrojový kód jednotlivých programů, případně přesný formát jednotlivých dat.

Součástí Digitálního archivu MU se na zpracování a uložení archiválií dále podílí mimo jiné tyto technologie:

7.1. Architektura hardware jednotlivých systémů je x86-64. Jednotlivé komponenty jsou tvořeny snadno zaměnitelnými servery a procesory různých výrobců, typicky servery komoditního typu, aby bylo možné průběžnou obnovu provádět nezávisle na konkrétním dodavateli, servery snadno zaměnit a případně opravit.

7.2. Operační systémy jednotlivých serverů jsou otevřeného linuxového typu v různých verzích dle životního cyklu jednotlivých serverů. Výjimkou jsou Konverzní a Antivirové servery, které mohou být hostovány na proprietárních operačních systémech jednotlivých formátů – pro formáty Microsoft Office se jedná o servery s operačním systémem Microsoft Windows, pro formáty Google Documents je konverze zajištěna prostřednictvím webové služby společnosti Google.

7.3. Souborové systémy pro uložení aktivních dat i jejich archivaci jsou s otevřeným formátem (XFS).

7.4. Databázový server pro uložení metadat Aktivní zóny je proprietární systém Oracle Database.

7.5. Webové servery jsou postaveny na otevřené technologii Apache, šifrování provozu webových aplikací na technologii OpenSSL.

7.6. Aplikační vrstvu tvoří vlastní aplikační prostředí v otevřeném programovacím jazyce Perl včetně kompilátorů a interpretů s otevřeným zdrojovým kódem.
7.7. Samotné aplikace jsou programovány na míru přímo Masarykovou univerzitou, tedy jsou mimo jiné k dispozici včetně zdrojového kódu.
7.8. Antivirové a konverzní aplikace jsou provozovány pod licencí třetích stran, typicky přímo výrobců jednotlivých proprietárních formátů, pro zajištění maximální shody převáděného a převedeného formátu.

Technologické zázemí vychází a je začleněno do širší infrastruktury Informačního systému Masarykovy univerzity, jehož dalších služeb také využívá. To se týká například e-mailových serverů pro zasílání automatických zpráv o provozu Digitálního archivu MU archivářům, správcům a případně ostatním uživatelům. Dále se jedná například o infrastrukturu pracovních stanic jednotlivých osob, které vyžadují výhradně standardní webový prohlížeč běžně dostupný pro danou platformu. Součástí technologií je také přímé napojení na systém elektronické spisové služby MU provozovaný a vyvíjený také v rámci Informačního systému MU, jehož součástí je rovněž přímo napojený systém elektronických spisoven.