Automatizovaná archivace kvalifikačních vysokoškolských prací se systémem Archivematica

Mark Jordan
Simon Fraser University 8888
University Drive
Burnaby, British Columbia, Canada
1-778-782-5753
mjordan@sfu.ca

ANOTACE

Tento článek popisuje nástroje, služby a pracovní postupy, které Simon Fraser University používá k automatizování přenosu jejich ETDs (pozn. Electronic Theses and Dissertations, kvalifikační vysokoškolské práce v elektronické podobě) z jejich uživatelského systému Thesis Registration System do systému Archivematica, platformy pro archivaci digitálních dat. Tento článek rovněž popisuje plány SFU na rozšíření jejich služeb digitální archivace prostřednictvím systému Archivematica, včetně integrace LOCKSS jako distribuovaného úložiště pro obsah spravovaný tímto systémem.

Kategorie a deskriptory

H.3.4 [Úložiště informací a vyhledávání]: systémy a software – distribuované systémy a H.3.7 [Úložiště informací a vyhledávání]: Digitální knihovny – Standardy

Obecné termíny

Management, Standardizace

Klíčová slova

Případové studie, digitální archivace, ETDs, pracovní postupy, automatizace, mikroslužby, OAIS, Drupal, Archivematica, LOCKSS

1. ÚVOD

Simon Fraser University (dále SFU) přijímá od studentů diplomové a dizertační práce a zprávy o absolventských projektech v digitální podobě od roku 2004. Ke konci roku 2012 knihovna zahájila provoz sady mikroslužeb sloužících k přenosu kvalifikačních vysokoškolských prací v elektronické podobě (dále ETDs) z jejího systému Theses Registration System (TRS)[1] do institucionálního

repozitáře Summit[2] bez lidského zásahu, vyjímaje zaměstnance knihovny, kteří potvrzují, že jsou práce připraveny k zveřejnění. Krátce po zahájení tohoto automatizovaného pracovního postupu knihovna začala přesouvat práce z TRS do systému pro archivaci digitálních dat Archivematica[3]. Tento proces je rovněž plně automatizován.

Tento článek popisuje zdůvodnění automatického přesunu ETDs do systému Archivematica, různé nástroje a služby, které se pro tuto automatizaci používají a také jejich společné fungování. Popisuje oblasti aktivního rozvoje, ve kterých SFU knihovna usiluje o rozšíření služeb digitální archivace.

2. CÍLE A ZÁKLADNÍ PRINCIPY

Kvalifikační vysokoškolské práce jsou jedním z nejdůležitějších typů vědeckých prací vytvářených na univerzitách. Ačkoli jsou kopie ETDs často distribuovány v komerčních službách jako je Proquest Dissertation Publishing[4] nebo v národních úložištích jako je Theses Canada[5], mnoho vzdělávacích institucí, které ETDs produkují, nesou za dlouhodobou archivaci těchto prací odpovědnost. Tento závazek však v průběhu času vyžaduje značné finanční prostředky.

Simon Fraser University se rozhodla jednat odpovědně podle svých závazků, ale dělá tak s cílem co možná největšího snížení nákladů. Mnoho nákladů spojených s digitální archivací je těžké předvídat[6], ale jedním aspektem této činnosti, ve kterém je možné relativně lehce snížit náklady, je lidská práce. K tomuto cíli SFU spěje prostřednictvím co možná nejkompletnější automatizace co největšího množství aspektů procesu zpracování ETDs. K vývoji služeb a procesu za účelem dosažení tohoto cíle vedou tři základní principy.

Prvním z nich je, že by archivace ETDs měla dodržovat spolehlivé osvědčené postupy založené na standardech digitální archivace jako například splnění požadavků referenčního modelu OAIS[7], používání PREMIS[8] archivace metadat, používání balící metody BagIt[9] a podpora standardních deskriptivních metadat jako například Dublin Core Terms.

Druhý základní princip zní, že všechny procesy související s archivací ETDs, které lze zautomatizovat, by měly být automatizovány. Lidský zásah bude v určitých bodech archivace potřebný, nicméně lidské zásahy a lokalizovaná rozhodnutí by měla být snížena na minimum.

Třetí princip - jakákoli speciální služba nebo zařízení používané v těchto procesech by mělo být jednoduše vyměnitelné. Z dlouhodobého hlediska se nástroje, považované za nejlepší ve své třídě, neustále mění. Je důležité, aby jakýkoli nástroj, který zlepšuje proces nebo provádí stejný proces za nižší náklady, mohl nahradit stávající nástroj v případě, že to příliš nenarušuje jiné procesy, které závisí na stávajícím nástroji. Kromě toho, schopnost nahradit služby a nástroje zjednodušuje přizpůsobení zbývajících prvků ostatním procesům digitální archivace.

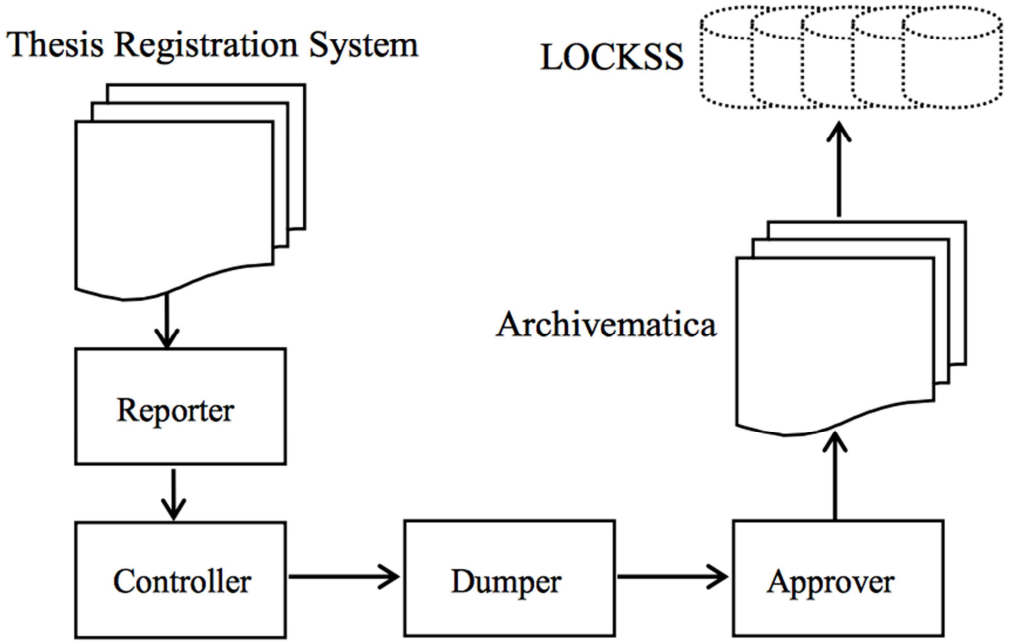
Tyto tři principy formují strukturu průběhu archivace popsanou níže.

Je důležité si povšimnout, že ETD není jen jednoduchý textový dokument. Mnoho ETDs obsahuje nezpracovaná data nebo údaje aplikací, multimediální obsah nebo další textové dokumenty, které s nimi souvisí. Tento dodatečný obsah se obvykle označuje jako „příloha“. ETDs navíc obvykle mají alespoň jeden metadatový popis označující název, datum zkompletování, obsah a tak dále, obvykle vyjádřený v ETD-MS[10] sadě. Archivace ETDs tudíž není tak jednoduchá jako zajistit uložení

kvalifikační práce v jediném PDF souboru. Dlouhodobá archivace ETDs musí brát v úvahu všechny výše uvedené typy obsahu[11].

3. STRUKTURA

Struktura archivace ETD se v Simon Fraser University skládá ze tří hlavních složek:
1) jejich systému Thesis Registration System, 2) sady mikroslužeb 3) systému digitální archivace Archivematica. Čtvrtá složka, Private LOCKSS Network, je v současné době ve vývoji. Následuje vizuální přehled struktury:



3.1. Thesis Registration System

Thesis Registration System umožňuje studentům jejich práci přihlásit a nahrát jakékoli související soubory do tzv. „odevzdávárny“. Jakmile student práci odevzdá, zaměstnanci knihovny ji předtím, než ji schválí ke zveřejnění v institucionálním repozitáři univerzity, překontrolují. Tento proces zahrnuje prověření, zda kvalifikační práce dodržují standardy pro zveřejňování prací nastavené univerzitou a všechna oprávnění pro zveřejnění byla studentem přijata.

Jakmile jsou všechny náležitosti kontroly splněny, zaměstnanci knihovny zaznamenají toto rozhodnutí v rámci odevzdání záznamu k práci tak, že jednoduše nadepíší políčko „připraveno ke zveřejnění“. Tento prvek odevzdávání prací je později používán v dotazu, který je spuštěn v noci za účelem identifikace všech odevzdaných prací, které byly schváleny ke zveřejnění během předchozího dne.
Systém Thesis Registration System je vybudovaný prostřednictvím systému Drupal[12], open-source systému pro správu obsahu. Drupal spravuje uživatelské účty a oprávnění, poskytuje mechanismus pro strukturu odevzdání prací a zachází s různými typy souborů, které musí studenti nahrát.

Zákaznický modul Drupal vyvinutý zaměstnanci knuhovny SFU řídí pracovní postupy zahrnuté do procesu prověřování odevzdaných prací a odesílá prostřednictvím e-mailu studentům zprávy, jakmile zaměstnanci knihovny provedou konkrétní kroky nebo učiní konkrétní rozhodnutí. Každé odevzdání vytvoří v systému Thesis Submission System instanci „node“, základní obsah struktury v systému Drupal.

3.2. Mikroslužby

Přenos obsahu ETDs ze systému Thesis Registration System do systému Achivematica je prováděn prostřednictvím malé řady mikroslužeb. Každá mikroslužba je shellový nebo PHP skript, který provádí jeden úkol nebo jednu skupinu souvisejících úkolů.

První mikroslužba (nazývaná „reporter“) se dotazuje na všechna odevzdání, která byla schválena předchozí den. Je to skript, který provádí dotazy popsané níže v části 3.1. Skript vypíše node IDs systému Drupal (který slouží jako jedinečný identifikátor každé odevzdané práce v systému Thesis Registration System) do datového souboru s aktuálním datem, které je zakódováno v jeho názvu.

Druhá mikroslužba („controller“) vytváří dva balící skripty na konkrétní úkol; jinými slovy, každý ze skriptů je spuštěn v systému. Tento přístup umožňuje spolehlivé ovládání chyb v každém skriptu a také umožňuje jednoduchou likvidaci dočasných souborů vytvořených balícími skripty. Regulátor je nastavený tak, aby se spustil každý den po spuštění mikroslužby „reporter“ a aby používal datové soubory vytvořené touto mikroslužbou jako vstupní data. V důsledku controller udělá smyčku po všech odevzdáních node IDs v rámci datového souboru pro aktuální den a spustí mikroslužby, “dumper” a “approver” v rámci příslušného odevzdání každého node ID.

Mikroslužba „dumper“ pokládá node ID odevzdání za parametr, dotazuje se systému Thesis Registration System na odpovídající node odevzdání a vytváří Dublin Core a ETD-MS soubory popisných metadat pro kvalifikační práce prostřednictvím informací ze záznamu odevzdání. Mikroslužba „dumper“ navíc určuje, které soubory jsou spojené s prací (kvalifikační práce v PDF, jakékoli přílohy, konkrétní oprávnění a další administrativní dokumenty) a také je zapisuje na disk. Na závěr „dumper“ zajistí, aby všechny soubory byly uspořádané ve struktuře podadresáře v souladu s metodou přenosu systému Archivematica (popsanému v následující části) a vytváří složku Bag, která obsahuje všechny soubory odevzdání.

Poslední mikroslužbou je „approver“, která kopíruje složku Bag vytvořenou mikroslužbou „dumper“ do serveru systému Archivematica a poté potvrzuje, že byl tento soubor úspěšně zkopírován a vydává HTTP požadavek systému Archivematica na schválení přenosu API (také popsaném v následující části).

3.3. Archivematica

Archivematica je open-source platforma pro archivaci digitálních dat. Standardizuje soubory do formátů vhodných pro archivaci prostřednictvím tzv. “formátové strategie”[13], a ukládá obsah do Archival Information Packages (AIPs) kompatibilních s OAIS. Archivematica propojuje množství open-source nástrojů jako je FITS[14], OpenOffice[15], FFmpeg[16] a Clam Antivirus[17] prostřednictvím své vlastní interní struktury mikroslužeb a využívá otevřené standardizované formáty jako je METS,[18] PREMIS, a BagIt k zajištění dlouhodobého řízení, které je založeno na standardech a přístupu k obsahu a k metadatům uložených v AIP balíčcích, které produkuje.

Obsah je do systému Archivematica přijímán jako „přenos“, což zahrnuje soubory k zachování, metadata popisující tyto soubory, dokumentaci odevzdání (oprávnění a další administrativní dokumenty) a popřípadě „konfigurační soubor“. Přenos strukturuje obsah pro přípravu na opětovné zabalení do OAIS Submission Information Package (SIP) a poté do Archival Information Package (AIP) pro dlouhodobou správu. Pokud má být obsah k dispozici uživatelské komunitě, povolí Archivematica za tímto účelem tvorbu Dissemination Information Packages (DIP).

Uživatelské rozhraní systému Archivematica specifikuje pracovní průběh zpracovávání dané sady souborů z přenosu do SIP, AIP, DIP do řady strukturovaných úkolů, z nich většina jsou zevnitř vytvořené instance jako mikroslužby. V rámci každé skupiny úkolů musí obslužný pracovník učinit několik rozhodnutí, například jestli standardizovat příchozí soubory pro archivaci, přístup (nebo oboje), zda schválit výsledky standardizace nebo ne, jestli k přenosu zažádat o další popisná metadata a kam uložit AIP. Jak jsou konkrétní typy souborů standardizované, určuje formátová strategie; například pro audio soubory může určit, že by měly být pro archivaci standardizovány do WAV formátu a pro přístup koncového uživatele do formátu MP3.

Pracovní postup může být automatizován prostřednictvím konfiguračního souboru, který zakóduje do strojově čitelného formátu každé rozhodnutí, které obslužný pracovník učiní, pokud manuálně zpracovával přenos. Schopnost automatizace rozhodování pracovního postupu je užitečná v případě, že systém Archivematica v dávkách zpracovává velké množství podobných přenosů nebo v případě, že lokální strategie určí, že dané rozhodnutí pracovního postupu by mělo platit vždy.

Za účelem zpracování ETDs od SFU konfigurační soubory stanovují, že by soubory měly být standardizovány jen pro archivaci (vzhledem k tomu nežádáme systém Archivematica o generování Dissemination Information Packages). Rovněž stanovují, který nástroj určující formát by měla Archivematica použít a kam uložit AIP.

Konfigurační soubor pouze odstraňuje potřebu obslužného pracovníka po přijetí přenosu systémem Archivematica. K automatizaci samotného příjmu Archivematica poskytuje REST API[19] pro schválení přenosů. Vzhledem k tomu, že API používá REST, je s API možná spolupráce v rámci skriptu spuštěného na jiném serveru (v tomto případě mikroslužba „approver“ spuštěná na server hostingu Thesis Submission System).

Je to kombinace této REST API a konfiguračního souboru, který umožňuje úplnou automatizaci přenosu obsahu ze zdroje jako je SFU Thesis Registration System do systému Archivematica, a poté prostřednictvím mikroslužeb digitální archivace systému Archivematica vytvoří OAIS-compliant Archival Information Package. V případě struktury SFU pro archivaci ETDs je tento proces vytvoření instance v dříve popsaných mikroslužnách „dumper“ a „approver“, které slučují a předávají obsah ETD do interních mikroslužeb systému Archivematica, jak jsou definovány konfiguračním souborem.

3.4. Dlouhodobá správa ETDs

V průběhu mohou být Archival Information Packages znovu získány i znovu přijaty do systému Archivematica jako SIPs (Submission Information Packages), kdy musí být obsah aktualizován nebo přesunut do nového formátu. Potřeba aktualizovat ETDs po zveřejnění je výjimečná, ale může se to stát a Faculty of Graduate Studies SFU má pro tyto případy zavedenou strategii.

Systém Archivematica podporuje autentičnost archivovaného obsahu tím, že kromě všech standardizovaných verzí vytvořených mikroslužbami (nebo externí standardizací do systému Archivematica) ukládá všechny originální dokumenty zahrnuté do přenosu. Systém také vytváří a ukládá kontrolní součty pro všechny soubory, aby v průběhu umožnil kontrolu a ověření bitové integrity. Na závěr jsou v implementaci SFU archivována všechna oprávnění autora ETDs ve stejném Archival Information Package jako ETDs dokument a přílohy doplněné o kontrolní součty.

3.5. Veřejný přístup do ETD

Tato verze ETD obsahu, která je systémem Archivematica přeměněna na OAIS Archival Information Package (AIP balíček), není určená pro přístup koncovému uživateli. AIP v podstatě obsahuje oprávnění a další citlivé informace, ke kterým by neměl mít koncový uživatel přístup.

V implementaci SFU jsou ETD a jejich související metadata přenášena přímo ze systému Thesis Registration System do institucionálního repozitáře Univerzity Summit pro veřejný přístup. Tento přenos je automatický a probíhá ve stejný čas jako přenos ETDs ze systému Thesis Registration System do systému Archivematica. Ve výsledku tedy probíhají dva procesy paralelně.
V institucionálním repozitáři mají koncoví uživatelé přístup k pracem prostřednictvím široké škály discovery nástrojů, jako je například jednotný discovery přístup a vyhledávací schopnosti samotného Summitu.

Systém Archivematica dobře zvládá tvorbu OAIS Dissemination Information Package (DIP) a přenos DIP do celé škály veřejně přístupných systémů správy obsahu a repozitářů včetně AtoM, CONTENTdm a DSpace. Implementace SFU tento prvek nepoužívá, protože proces přenosu ETDs ze systému Thesis Registration System do Summitu už byl v provozu, když knihovna začala používat systém Archivematica. Bylo by možné vytvořit nové mikroslužby systému Archivematica za účelem vytvoření DIP pro Summit SFU, nicméně knihovna zvolila alternativní přístup k integraci systému Archivematica a jeho institucionálního repozitáře popsaný níže v části 4.3.

4. Vývoj podrobného plánu

Knihovna SFU aktivně pracuje na rozšíření integrace současných ETDs archivačních služeb s několika dalšími nástroji.

4.1. Integrace LOCKSS

V současné době se pracuje na tom, aby systém Archivematica mohl ukládat AIP balíčky do Private LOCKSS Network (PLN)[20]. Tento rozvoj umožní automatický přesun AIP a jeho sklizení pomocí LOCKSS a začlenění do PLN. Ukládání AIPs do Private LOCKSS Network zajistí, že budou identické kopie spravovány v distribuovaném zajištěném prostředí. Knihovna SFU a skupina partnerských institucí úzce spolupracují s vývojáři systému Archivematica, aby zajistili, že tato práce bude v souladu s novým ukládáním API, které se pro Archivematicu vyvíjí. Tato API umožní použití široké škály platforem pro ukládání AIPs, které vytváří.

4.2. Akademické přezkoumání kvalifikačních prací prostřednictvím Open Journal Systems

Ačkoli to přímo nesouvisí s archivací ETDs, Faculty of Graduate Studies na SFU plánuje používání Open Journal Systems (OJS)[21] pro akademické přezkoumání prací. Open Journal Systems poskytují sadu nástrojů pro odevzdání rukopisu, vzájemné hodnocení a redakční pracovní postup pro časopisecké články, které se snadno přizpůsobí přezkoumání prací akademickou komisí. Knihovna SFU bude úzce spolupracovat s fakultou, aby bylo zajištěno, že se ETDs souvisle přesune z OJS do systému knihovny Thesis Submission System a odtud dále prostřednictvím struktury digitální archivace popsané v tomto článku.

4.3. Automatická archivace obsahu institucionálního repozitáře SFU

Nástroje a pracovní postupy popsané v tomto článku mohou být také aplikovány v automatické archivaci obsahu odevzdaného do institucionálního repozitáře SFU Summit. V současné době probíhá implementace takového procesu. Vše, co je zapotřebí, je v podstatě pozměnění mikroslužby „dumper“ tak, aby převedl i jiné položky než ETDs v institucionálním repozitáři do složky přenosu systému Archivematica. Jiné položky než ETDs v repozitáři SFU jsou předtisky časopisů a kapitol knih, dokumenty z konferencí, hlášení a další práce podávané přímo koncovým uživatelem a zaměstnanci knihovny jako služba univerzitní komunitě. Automatický přesun obsahu ze Summitu do systému Archivematica bude poskytován službami spolehlivé digitální archivace, které v mnoha institucionálních repozitářích chybí.

Schopnost nahradit jednu část struktury digitální archivace SFU (Thesis Registration System) jinou (institucionální repozitář) a provést menší změny jedné mikroslužby („dumper“) ukazuje důležitý základní princip struktury: „jakákoli konkrétní služba nebo nástroj použitý v procesu by měl být jednoduše vyměnitelný“. Tento systém může být aplikován i na ostatní zdroje obsahu knihovny SFU, které potřebují archivovat, jako jsou místní digitalizované rukopisné sbírky a noviny, data výzkumu a archivované webové stránky.

5. ODKAZY

[1] Simon Fraser University’s Thesis Registration System. https://theses.lib.sfu.ca

[2] Summit, Simon Fraser University’s Institutional Repository. http://summit.sfu.ca

[3] Archivematica. https://archivematica.org

[4] Proquest Dissertation Publishing. http://www.proquest.com/en-US/products/dissertations/

[5] Theses Canada. http://www.collectionscanada.gc.ca/thesescanada/index- e.html

[6] Wheatley, Paul. 2012. Digital Preservation Cost Modelling: Where did it all go wrong? Blog post. http://openplanetsfoundation.org/blogs/2012-06-29-digital- preservation-cost-modelling-where-did-it-all-go-wrong

[7] Reference Model For An Open Archival Information System (OAIS). http://public.ccsds.org/publications/archive/650x0m2.pdf

[8] PREMIS Data Dictionary for Preservation Metadata. http://www.loc.gov/standards/premis/

[9] The BagIt File Packaging Format. http://tools.ietf.org/html/draft-kunze-bagit-09

[10] ETD-MS: an Interoperability Metadata Standard for Electronic Theses and Dissertations. http://www.ndltd.org/standards/metadata/

[11] Shreeves, Sarah L. 2013. Supplemental Files in Electronic Theses and Dissertations: Implications for Policy and Practice. Poster presented at the 8th International Digital Curation Conference, Amsterdam, Netherlands, January 14- 17, 2013. http://hdl.handle.net/2142/35314

[12] Drupal. http://drupal.org/

[13] Archivematica Format Policies. https://www.archivematica.org/wiki/Media_type_preservation_plans

[14] File Information Tool Set (FITS). http://code.google.com/p/fits/

[15] OpenOffice. http://www.openoffice.org/

[16] FFmpeg. http://www.ffmpeg.org/

[17] ClamAV. http://www.clamav.net/lang/en/

[18] Metadata Encoding and Transmission Standard (METS). http://www.loc.gov/standards/mets/

[19] Approving a transfer. https://www.archivematica.org/wiki/Administrator_manual_0.10#Approving_a_transfer

[20] Lots of Copies Keep Stuff Safe (LOCKSS). http://www.lockss.org/

[21] Open Journal Systems (OJS). http://pkp.sfu.ca/ojs