Univerzita Karlova v Praze
Filozofická fakulta
Ústav informačních studií a knihovnictví

Studijní program: Informační studia a knihovnictví
Studijní obor: Informační věda, kombinovaná forma studia

Ing. Jan Mach

DISERTAČNÍ PRÁCE

Správa, vyhledávání a zpřístupňování elektronických vysokoškolských kvalifikačních prací

Management, Retrieval and Access to Electronic Theses and Dissertations

Školitelka PhDr. Eva Bratková, Ph.D.

2015

Poděkování

Rád bych především poděkoval vedoucí disertační práce PhDr. Evě Bratkové, Ph.D. za cenné připomínky a podporu při psaní této disertační práce i za spolupráci při řešení problematiky zpřístupňování vysokoškolských kvalifikačních prací v ČR.

Mé poděkování patří také všem respondentkám a respondentům, bez jejich otevřenosti a ochoty by nebylo možné zpracovat průzkum zpřístupňování vysokoškolských kvalifikačních prací v takovémto rozsahu.

Velké poděkování za trpělivost a podporu patří mé rodině a přátelům.

Prohlášení

Prohlašuji, že jsem disertační práci vypracoval samostatně, že jsem řádně citoval všechny použité prameny a literaturu, a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

V Praze dne 24. března 2015 _____

podpis

Abstrakt (CZ)

Disertační práce je věnována analýze současné praxe a trendům provozu repozitářů elektronických vysokoškolských kvalifikačních prací (eVŠKP) z pohledu jejich správy, vyhledávání a zpřístupňování. První část představuje výchozí pojmy a současný stav zpřístupňování eVŠKP v českých a zahraničních repozitářích, obsahuje průzkum zpřístupňování eVŠKP v ČR z roku 2014, kterého se zúčastnily všechny veřejné vysoké školy. V druhé části práce je představen metadatový soubor EVSKP-MS, konkrétně možnosti mapování prvků EVSKP-MS na další metadatové formáty a využití standardu v rámci OAI-PMH protokolu. Problematika zpřístupňování eVŠKP je dále řešena z pohledu vhodnosti metrik pro evaluaci využití distribuovaných eVŠKP. Vyhledáváním eVŠKP se zabývají popsané případové studie a doporučení pro výběr discovery služby a pro tvorbu vyhledávacího serveru metadat eVŠKP a související uživatelské rozhraní s fasetovým vyhledáváním. Poslední část disertační práce se zabývá problematikou plagiátorství. V práci je představena analýza nejvýznamnějších systémů na podporu vyhledávání plagiátů a vývoj portálu Validátor VŠE pro zpřístupnění výsledků kontroly dokumentů.

Klíčová slova (CZ)

vysokoškolské kvalifikační práce, EVSKP-MS, repozitáře, OAI-PMH, metadata, fasety, plagiátorství

Abstract (EN)

The dissertation is devoted to analysis of current practice and trends in providing repositories of electronic theses and dissertation (ETDs) in terms of their management, searching and dissemination. The first part presents terminology and the current state of access to ETDs in Czech and foreign repositories and includes results of a survey of the state of access to ETDs in the Czech Republic which was completed in 2014 by all public universities. In the second part, a metadata standard is presented, particularly the possibility of mapping EVSKP-MS metadata elements to other metadata formats and utilization within the OAI-PMH protocol. The issue of access to ETDs is dealt with further in terms of metrics for an evaluation of usage of distributed ETDs. Searching for ETDs is also described in case studies as are recommendations for public tenders for a discovery service and for creating an ETD metadata search server and an associated user interface with faceted search. The final part of the thesis focuses on the issue of plagiarism. This incorporates a presentation and analysis of the most important plagiarism detection systems and a case study of the development of the portal Validátor VŠE to provide access to results of document analysis.

Keywords (EN)

ETD, EVSKP-MS, repositories, OAI-PMH, metadata, facets, plagiarism

Obsah

Seznam zkratek

Seznam obsahuje v textu často používané zkratky, rozepsané a případně doplněné o URL adresu s dalšími informacemi.

AKVŠ	Asociace knihoven vysokých škol ČR http://www.akvs.cz
AMU	Akademie múzických umění v Praze http://www.amu.cz
API	aplikační programové rozhraní (angl. application programming interface)
CIKS	Centrum informačních a knihovnických služeb VŠE v Praze http://ciks.vse.cz
CorpCZ	Metadatový soubor pro popis korporací verze 1.0 http://www.evskp.cz/standardy/corpcz/
COUNTER	Counting Online Usage of Networked Electronic Resources http://www.projectcounter.org
DC	metadatové formáty Dublin Core Metadata Iniciative (viz DC Set a DCMI Terms)
DC Set	Dublin Core Metadata Element Set (ISO Standard 15836:2009, ANSI/NISO Standard Z39.85-2012) http://dublincore.org/documents/dces/
DCMI	Dublin Core Metadata Iniciative http://dublincore.org
DCMI Terms	Dublin Core Metadata Iniciative Metadata Terms http://dublincore.org/documents/dcmi-terms/
DEEP	DART-Europe E-theses Portal http://www.dart-europe.eu
EIZ	Elektronické informační zdroje
ETD-MS	An Interoperability Metadata Standard for Electronic Theses and Dissertations http://www.ndltd.org/standards/metadata
EThOS	the Electronic Theses Online System http://ethos.bl.uk
EVSKP-MS	Metadatový soubor pro elektronické vysokoškolské kvalifikační práce v ČR http://www.evskp.cz/standardy/metadata/
eVŠKP	elektronické vysokoškolské kvalifikační práce (angl. Electronic Theses and Dissertations, ETDs)
GPL	software GooglePlagiarism vyvinutý autorem disertační práce
JIF	impakt faktor časopisu (angl. Journal Impact Factor)
JISC	Joint Information Systems Committee http://www.jisc.ac.uk
Komise eVŠKP	Odborná komise pro otázky elektronického zpřístupňování vysokoškolských kvalifikačních prací AKVŠ ČR http://www.evskp.cz
MARC	Strojově čitelná katalogizace (angl. MAchine Readable Cataloging)

MUNI	Masarykova univerzita http://www.muni.cz
MŠMT ČR	Ministersto školství, mládeže a tělovýchovy ČR
NDLTD	Networked Digital Library of Theses and Dissertations http://www.ndltd.org
NTK	Národní technická knihovna http://www.techlib.cz
NUŠL	Národní úložiště šedé literatury http://www.nusl.cz, http://nusl.techlib.cz
OAI-PMH	The Open Archives Initiative Protocol for Metadata Harvesting http://www.openarchives.org/OAI/openarchivesprotocol.html
PersCZ	Metadatový soubor pro popis fyzických osob http://www.evskp.cz/standardy/perscz/
PIRUS	Publisher and Institutional Repository Statistics
PLG	textový korpus použitý k testování (100% plagiát)
PQDT	databáze ProQuest Dissertations Abstracts International
PTS	metadatový formát Theses.cz http://theses.cz/pts/elements/1.0/
SKIP ČR	Svaz knihovníků a informačních pracovníků České republiky http://www.skipcr.cz
THE, Theses.cz	národní registr vysokoškolských kvalifikačních prací a systém na kontrolu plagiátů http://www.theses.cz
TUR	software Turnitin http://turnitin.com
UK	Univerzita Karlova v Praze
VŠ	vysoké školy (veřejné, soukromé a státní)
VŠE v Praze	Vysoká škola ekonomická v Praze http://www.vse.cz
VŠKP	Vysokoškolské kvalifikační práce (v elektronické nebo v tištěné verzi)
VVŠ	veřejné vysoké školy

Seznam ilustrací

Seznam tabulek

Seznam grafů

1 Úvod

Disertační práce *Správa, vyhledávání a zpřístupňování elektronických vysokoškolských kvalifikačních prací* je výsledkem osmiletého výzkumu autora v oblasti projektování a provozu digitálních repozitářů elektronických vysokoškolských kvalifikačních prací (zkráceně eVŠKP) v České republice.

1.1 Předmět, cíle a metody

Workflow vysokoškolských kvalifikačních prací v elektronické podobě zahrnuje mnoho dílčích předmětných oblastí, kterými jsou např. zadání tématu, formální úprava práce, vložení plného textu a příloh eVŠKP do informačního systému, metadatový popis, antiplagiátorská kontrola, indexování, vyhledávání a export eVŠKP, archivace aj.

Na základě analýzy stavu zpřístupňování eVŠKP v ČR a zkušeností s budováním a provozem digitálních repozitářů autor vybral tyto tři předmětné oblasti pro výzkum disertační práce:

- mapování metadatových prvků, komunikace a vyhledávání záznamů eVŠKP,
- metriky pro hodnocení užití záznamů eVŠKP v otevřených repozitářích,
- plagiátorství u eVŠKP.

Cílem disertační práce je poskytnout ověřená řešení problémů souvisejících s výše uvedenými oblastmi správy, vyhledávání a zpřístupňování eVŠKP. Autor ve své disertační práci hledá především řešení následujících problémů a souvisejících otázek:

1) Poslední velký průzkum provedla Komise eVŠKP AKVŠ ČR v roce 2009. Jaký je aktuální stav zpřístupňování eVŠKP v ČR?
2) Pro popis eVŠKP se používají různé metadatové formáty. Jak mapovat prvky standardu EVSKP-MS do jiných metadatových sad a záznamy následně exportovat do dalších repozitářů a služeb v ČR a v zahraničí?
3) Záznamy o jedné eVŠKP bývají dostupné v několika repozitářích zároveň. Jaké metriky jsou vhodné pro hodnocení užití záznamů eVŠKP v repozitářích s otevřeným přístupem a jak je měřit?
4) České repozitáře eVŠKP oproti zahraničním zaostávají v možnostech vyhledávání. Knihovny mají zájem o zlepšení vyhledávacích rozhraní, uživatelům nabízejí

např. tzv. discovery služby. Jak vyhledávat eVŠKP v institucionálním repozitáři, jak vybrat discovery službu, která by indexovala potřebné elektronické informační zdroje, katalog knihovny i repozitář eVŠKP v centrálním indexu poskytovatele?

5) Pro podporu antiplagiátorské kontroly eVŠKP potřebují vysoké školy vybrat vhodný nástroj a výsledky kontroly zpřístupnit akademické obci. Jaký antiplagiátorský nástroj vybrat a jak lze výsledky kontroly zpřístupnit odpovědným osobám?

Mezi metody použité při přípravě disertační práce patří především komparativní a kritická analýza publikovaných pramenů a konkrétních repozitářů eVŠKP v ČR a ve světě, vlastní výzkum s dotazníkovým šetřením, návrh a evaluace metadatových specifikací a implementace prototypů a modelových řešení. Navržená doporučení a postupy byly autorem experimentálně ověřovány v praxi a popsány ve formě případových studií.

Výsledky disertační práce tvoří popis realizovaných výzkumů, jejich analýza a zhodnocení. Konkrétní přínosy a význam pro rozvoj vědního oboru a pro implementaci postupů v praxi dále tvoří metodologická doporučení a popsané případové studie implementace aplikací. Jedná se především o autorem zpracované doporučení pro mapování prvků metadatového standardu EVSKP-MS při zpřístupňování eVŠKP, doporučení pro implementaci vyhledávání eVŠKP, případové studie vývoje aplikací *Vyhledávací rozhraní eVŠKP* a *Validátor VŠE*.

1.2 Stylistika textu práce, citování

Autor v textu používá vyjadřování ve třetí osobě jednotného čísla, případně trpného rodu. V případě, kdy se na výzkumu ve významné míře podílelo více osob, je tato skutečnost v textu konkrétně uvedena.

V textu je použito citování podle citační normy *ČSN ISO 690 (01 0197)* platné od 1. dubna 2011. Při odkazování na očíslovaný seznam literatury je v textu použito metody číselných odkazů v kulatých závorkách.

V textu pod čarou jsou uvedeny doplňující informace k textu. V případě, kdy obsah kapitoly vychází z autorova publikovaného textu, je na to upozorněno v úvodu kapitoly.

Příklady zdrojového kódu, značky XML, URL adresy a názvy souborů jsou zapsány fontem Microsoft Sans Serif v tmavě šedé barvě.

1.3 Struktura práce

V kapitole 2 disertační práce autor definuje základní používané termíny, analyzuje vývoj repozitářů eVŠKP, současný stav a trendy ve zpřístupňování eVŠKP v ČR. Komplexnějšímu porozumění řešené problematice přispívá průzkum zpřístupňování eVŠKP na veřejných vysokých školách v ČR z roku 2014 popsaný v kapitole 3. Provedený průzkum navazuje na předchozí dotazníková šetření, která se uskutečnila v letech 2006 – 2009 během působení autora v rámci Odborné komise pro otázky elektronického zpřístupňování vysokoškolských kvalifikačních prací (zkráceně Komise eVŠKP) při Asociaci knihoven vysokých škol ČR (zkráceně AKVŠ ČR).

Autor práce se přes deset let zabývá problematikou metadatového popisu vysokoškolských kvalifikačních prací v digitálních repozitářích. V rámci Komise eVŠKP byl, spolu s Evou Bratkovou, zpracovatelem *Standardizačního soubor metadatových prvků určených pro popis vysokoškolských kvalifikačních prací obhajovaných na vysokých školách v ČR a pro přenos souborů EVSKP-MS* (1) a souvisejících standardů pro popis fyzických osob PersCZ (2) a korporací CorpCZ (3). Tyto výsledky práce jsou detailněji popsány v disertační práci Evy Bratkové.

Autor na popis metadatových standardů v této disertační práci navazuje. V kapitole 4 *Mapování metadat eVŠKP* je představeno vypracované doporučení pro mapování prvků standardu EVSKP-MS do dalších vybraných metadatových sad na základě komparační analýzy jednotlivých metadatových formátů. Sjednocení mapování prvků umožňuje jednoznačný převod formátů potřebný pro usnadnění exportu metadat eVŠKP do repozitářů v ČR a v zahraničí.

Případová studie, popsaná v podkapitole 4.3, ukazuje využití doporučeného mapování na příkladu implementace Open Archives Initiative Protocol for Metadata Harvesting (zkráceně OAI-PMH) nad daty Databáze kvalifikačních prací VŠE, s exportem metadat do národního repozitáře VŠKP – Theses.cz (zkráceně Theses.cz), Národního úložiště šedé literatury (zkráceně NUŠL) a evropského repozitáře DART-Europe E-theses Portal (zkráceně DEEP).

V souladu s trendem otevřeného přístupu (viz *Definice Open Access* v oddílu 2.1.4) dochází k rozšiřování metadat a plných textů eVŠKP do národních i mezinárodních repozitářů. Měřit

užití plného textu v elektronické podobě v jednom repozitáři není pro správce systémů obtížné. Otázkou však zůstává, jak měřit užití eVŠKP v případě sdílení metadat a plných textů v online prostředí Internetu. Kapitola 5 analyzuje metriky pro tištěné a online dostupné publikace, altmetriky založené na analýze sociálních sítí a projekty měření impaktu u publikací s otevřeným přístupem. Vhodnost altmetrik a webometrik autor analyzuje na základě dat repozitáře Pittsburské univerzity. V závěru kapitoly je formulováno doporučení pro výpočet metrik užití eVŠKP v repozitářích ČR.

Kapitoly 6 a 7 se zabývají otázkou vyhledávání vysokoškolských kvalifikačních prací. Případová studie *Výběr systému centralizovaného vyhledávání* řeší aktuální problém vysokoškolských knihoven s výběrem vhodné discovery služby – webové aplikace umožňující pomocí jednoho vyhledávacího rozhraní prohledávat metadata a plné texty z elektronických informačních zdrojů, katalogu knihovny a repozitářů univerzity, včetně metadat a plných textů vysokoškolských kvalifikačních prací. Nevhodně formulované podmínky mohou vést k nutnosti zrušit a opakovat výběrové řízení, jak tomu bylo např. v případě zrušeného výběrového řízení Univerzity Pardubice *Dodávka discovery systému* z roku 2011. Vhodná formulace zadávací dokumentace je prezentována na příkladu výběrového řízení Univerzity J. E. Purkyně v Ústí nad Labem, pro kterou autor disertace zpracovával podklady k veřejné zakázce *Systém centralizovaného vyhledávání elektronických informačních zdrojů* (4) a účastnil se hodnocení nabídek jako přizvaný specialista.

Kapitola 7 *Vyhledávací rozhraní eVŠKP* vychází z analýzy uživatelských rozhraní zahraničních a českých repozitářů eVŠKP (v podkapitole 2.3) a metadat eVŠKP ve formátu EVSKP-MS (v kapitole 4). Autor disetační práce na příkladu modelové aplikace řeší redesign vyhledávacího rozhraní repozitáře eVŠKP za využití fasetového vyhledávání. Pro indexování a vyhledávání metadat je použita open source platforma Apache Solr, v rámci které je autorem realizován import, parsování metadat ve standardu EVSKP-MS a vyhledávání nad vytvořeným indexem.

Poslední část výzkumu v disertační práci je věnována plagiátorství. Problematika opisování je autorovi známa mj. z práce pro Českou televizi, pro kterou zpracovával analýzy plagiátorství u vysokoškolských kvalifikačních prací obhajovaných např. na Fakultě právnické Západočeské univerzity v Plzni aj. (viz *Publikační činnost autora* uvedená v závěru práce).

Pro podporu vyhledávání duplicit je dostupných několik nástrojů, ty nejvyužívanější pro odhalování plagiátorství u eVŠKP v ČR byly autorem analyzovány v rozsáhlém testu popsaném v kapitole 8. Na prototypovém návrhu vlastní aplikace autor ukazuje potenciál vyhledávání duplicit eVŠKP vůči zdrojům na Internetu za využití vyhledávače Google. Výsledky testu, prezentované odborné veřejnosti v rámci 6. ročníku Semináře ke zpřístupňování šedé literatury, jsou využitelné nejen při výběru vhodného nástroje pro univerzitu, ale jsou i námětem pro další rozvoj Theses.cz a dalších služeb.

Jak ukázal průzkum zpřístupňování vysokoškolských kvalifikačních prací v roce 2014, pouze 15 veřejných vysokých škol oponentům zpřístupňuje ve svém informačním systému výsledky kontroly eVŠKP ze systému Theses.cz. Kapitola 9 formou případové studie popisuje návrh a vývoj aplikace Validátor VŠE, která je uživatelským rozhraním mezi repozitáři VŠE v Praze a systémy Masarykovy univerzity v Brně (zkráceně MUNI). Aplikace zajišťuje jednoduché a přehledné zpřístupnění výsledků kontroly eVŠKP a dalších textů vyučujícím VŠE v Praze. Je tak modelovým řešením i pro další univerzity, které výsledky kontroly oponentům doposud nezpřístupňují.

Dílčí závěry a doporučení jsou formulovány na konci každé kapitoly.

1.4 Odborná základna pro disertační práci

V lednu 2004 se autor disertační práce spolupodílel na založení Odborné komise pro otázky elektronického zpřístupňování vysokoškolských kvalifikačních prací pod Asociací knihoven vysokých škol ČR. Do roku 2008 působil v komisi jako řadový člen, od roku 2008 se po Ivě Horové stal předsedou. V rámci komise intenzivně spolupracoval především s odbornou garantkou a místopředsedkyní komise Evou Bratkovou.

Cíle komise byly:

- analýza významných zahraničních systémů eVŠKP,
- analýza českých lokálních systémů eVŠKP,
- projektování digitální knihovny eVŠKP na národní úrovni,
- budování a provoz národního digitálního systému eVŠKP,
- mezinárodní kooperace v oblasti zpřístupňování eVŠKP.

Během své práce v Komisi eVŠKP autor získal velmi cenné praktické zkušenosti s problematikou zavádění elektronických vysokoškolských kvalifikačních prací na českých vysokých školách, především při návrhu repozitářů eVŠKP a souvisejících metadatových standardů. Osobně se pak mj. aktivně podílel na přípravě doporučení pro české vysoké školy, na workflow odevzdávání eVŠKP, na návrhu projektu národního registru vysokoškolských kvalifikačních prací Theses.cz a následné spolupráci s provozovateli Theses.cz, na tvorbě standardů pro popis eVŠKP, osob a korporací. Získané zkušenosti a doporučení autor předával knihovnické komunitě ve svých přednáškách na Seminářích ke zpřístupňování eVŠKP (pořádaných v rámci Komise eVŠKP), na Seminářích ke zpřístupňování šedé literatury (které společně s Národní technickou knihovnou pořádal od roku 2008 v rámci projektu Národního úložiště šedé literatury) a dalších.

Od roku 2010 se problematika zpřístupňování eVŠKP v ČR rozšířila o nová témata, jako je Open Access, problematika plagiátorství, zpřístupňování českých vysokoškolských prací v zahraničních repozitářích, využití discovery služeb a optimalizace nástrojů pro vyhledávání eVŠKP. Autor se proto po ukončení činnosti Komise eVŠKP nadále profesně věnuje zpřístupňování eVŠKP a nově i zpřístupňování odborných vědeckých publikací. Vybrané výsledky práce doktoranda související s tématem disertační práce jsou detailněji představeny v jednotlivých kapitolách.

1.4.1 Související řešené projekty

Předkládaný výzkum je založen mj. na dlouholeté práci autora v knihovně VŠE – v Centru informačních a knihovnických služeb (zkráceně CIKS), kde pracuje jako vedoucí odboru Informační podpory studia a výzkumu. V rámci CIKS se podílel jako řešitel nebo spoluřešitel na řadě projektů, které souvisely s vývojem Databáze kvalifikačních prací VŠE, národním registrem VŠKP (Theses.cz), Národním úložištěm šedé literatury, se standardizací metadat

pro eVŠKP v ČR a kontrolou vysokoškolských kvalifikačních prací na duplicity – plagiátorstvím. Jedná se konkrétně o tyto projekty přímo související s disertační prací:

- *Dlouhodobé ukládání a archivace digitálních dokumentů dle zákona č. 499/2004 Sb.* (MŠMT ČR – centralizovaný rozvojový projekt C17, 2015, spoluřešitel CIKS Jan Mach)
- *Digitální knihovna pro šedou literaturu - funkční model a pilotní realizace* (MK ČR, 2007-2011, příjemce NTK v Praze, spolupříjemce VŠE v Praze, řešitel Jan Mach)
- *Meziuniverzitní síť technických a metodických opatření na ochranu proti plagiátorství* (MŠMT - centralizovaný rozvojový projekt C39, 2011, spoluřešitel CIKS Jan Mach)
- *Rozvoj infrastruktur pro využívání hledání podobností mezi studentskými pracemi a zdroji na Internetu* (MŠMT - centralizovaný rozvojový projekt C20, 2010, spoluřešitel CIKS Jan Mach)
- *Odhalování plagiátů v seminárních pracích* (MŠMT - centralizovaný rozvojový projekt C13, 2009, spoluřešitel CIKS Jan Mach)
- *Národní registr VŠKP a systém na odhalování plagiátů* (MŠMT - centralizovaný rozvojový projekt č. C1, 2008, řešitel VŠE v Praze Jan Mach)

Vybrané výsledky výše uvedených projektů jsou detailněji popsány v následujících kapitolách.

1.4.2 Členství v odborných orgánech

Autor disertační práce působil nebo stále působí v těchto odborných orgánech:

- Asociace knihoven vysokých škol ČR (AKVŠ ČR)
  2002 – současnost, konsorcionální členství VŠE v Praze,
  od roku 2015 členem výkonného výboru AKVŠ ČR
- DART-Europe Board
  2014 - současnost, člen rady
- European Business School Librarians' Group (EBSLG)
  1996 – současnost, konsorcionální členství CIKS,
  od roku 2014 osobně zástupce VŠE v Praze

- Iniciativa na podporu Open Access v ČR
  2010 – současnost, člen iniciativy, správce webu http://www.openaccess.cz
- Klub vysokoškolských knihovníků SKIP ČR
  kolektivní členství CIKS, správce webu Klubu vysokoškolských knihovníků SKIP
- Odborná komise pro otázky elektronického zpřístupňování VŠKP AKVŠ ČR
  2004-2010, od roku 2008 předseda komise
- Rada pro informatizaci VŠE v Praze
  2010 – současnost, člen rady
- Rada pro vývoj Národního úložiště šedé literatury (NUŠL)
  2013 – současnost, člen rady

2 Výchozí stav

V úvodu kapitoly jsou popsány základní pojmy, které se týkají zpřístupňování vysokoškolských kvalifikačních prací. Následující část obsahuje kritickou a komparativní analýzu stavu zpřístupňování vysokoškolských kvalifikačních prací do roku 2014 a souvisejících repozitářů.

2.1 Terminologie

Předmětem této disertační práce jsou vysokoškolské kvalifikační práce[1] (zkráceně VŠKP) a jejich zveřejňování v elektronické formě (zkráceně eVŠKP). Repozitáře eVŠKP mohou plnit roli zpřístupňování, ale i archivování prací. Samostatné části jsou proto věnovány definici VŠKP, autorskoprávní problematice a pojmu Open Access – trvalému, okamžitému a bezplatnému přístupu k výsledkům vědy a výzkumu, včetně eVŠKP, na Internetu.

2.1.1 Definice vysokoškolských kvalifikačních prací

Vysokoškolské kvalifikační práce jsou definovány *zákonem č. 111/1998 sb., o vysokých školách a o změně a doplnění dalších zákonů (Zákon o vysokých školách)* (5). Jedná se o práce, jejichž zpracováním student při obhajobě prokazuje nabyté zkušenosti během studia.

V případě bakalářského studijního programu se studium ukončuje státní závěrečnou zkouškou, jejíž součástí je zpravidla obhajoba bakalářské práce. V případě magisterského studijního programu se studium ukončuje státní závěrečnou zkouškou, jejíž součástí je obhajoba diplomové práce. V oblasti lékařství a veterinárního lékařství a hygieny se studium řádně ukončuje státní rigorózní zkouškou. Absolventi magisterských studijních programů, kteří získali akademický titul magistr, mohou vykonat v téže oblasti studia státní rigorózní zkoušku, jejíž součástí je obhajoba rigorózní práce. Doktorský studijní program se zakončuje doktorskou zkouškou a obhajobou disertační práce.

[1] Definice vysokoškolských kvalifikačních prací v textu kapitoly vychází z kapitoly *Zpřístupnění vysokoškolských kvalifikačních prací* publikace *Repozitáře šedé literatury* (4), autorem kapitoly je Jan Mach. Publikace vznikla v rámci řešení projektu *Digitální knihovna pro šedou literaturu – funkční model a pilotní realizace*, který podpořilo Ministerstvo kultury České republiky v rámci programových projektů.

Zatímco pro práce bakalářské, diplomové a rigorózní není v zákoně o vysokých školách přesněji stanoveno, jakou mají mít formu a obsah, v případě doktorské zkoušky a disertační práce musí student prokázat „schopnost a připravenost k samostatné činnosti v oblasti výzkumu nebo vývoje nebo k samostatné teoretické a tvůrčí umělecké činnosti. Disertační práce musí obsahovat původní a uveřejněné výsledky nebo výsledky přijaté k uveřejnění.“ (5 § 47).

Mezi práva studenta daná vysokoškolským zákonem patří právo navrhovat téma své bakalářské, diplomové, rigorózní nebo disertační práce. Témata VŠKP jsou dále vypisována podle zaměření jednotlivými vyučujícími vysoké školy a studenti si mohou téma z nabídky vybrat. Stejné téma může být zpracováváno i více studenty. Registry VŠKP jsou proto vhodným podkladem pro analýzu témat vedených nebo obhajovaných závěrečných prací, díky které můžeme stanovit bližší specializaci konkrétního vyučujícího.

Mezi vysokoškolské kvalifikační práce můžeme počítat i habilitační práce, které předkládají žadatelé při habilitačním řízení za účelem získání titulu docent. V tomto případě se nejedná o studenty předkládající práci v rámci studijního programu školy. „V habilitačním řízení se ověřuje vědecká nebo umělecká kvalifikace uchazeče, a to zejména na základě habilitační práce a její obhajoby a dalších vědeckých, odborných nebo uměleckých prací, a jeho pedagogická způsobilost na základě hodnocení habilitační přednášky a předcházející pedagogické praxe (…) Habilitační prací se rozumí:

a) písemná práce, která přináší nové vědecké poznatky, nebo
b) soubor uveřejněných vědeckých prací nebo inženýrských prací doplněný komentářem, nebo
c) tiskem vydaná monografie, která přináší nové vědecké poznatky, nebo
d) umělecké dílo nebo umělecký výkon nebo jejich soubor, kterým je například vynikající veřejná umělecká činnost.“ (5 § 72)

Vysokoškolské kvalifikační práce obsahují výsledky vědecké, výzkumné, vývojové nebo umělecké činnosti (především práce disertační a habilitační), procházejí obhajobou – recenzním řízením, nejsou většinou publikovány, a proto ani nejsou jednoduše dostupné. Tvoří tak významnou část šedé literatury a existuje zde oprávněný zájem na jejich zpřístupnění veřejnosti minimálně z důvodu transparentnosti udělování VŠ titulů.

Zpřístupněním šedé literatury, se zaměřením na VŠKP, se autor zabýval v rámci projektu Národního úložiště šedé literatury (viz oddíl 2.3.2, část *Národní úložiště šedé literatury*).

2.1.2 Vysokoškolské kvalifikační práce jako archiválie

Naskýtá se otázka, zda vysokoškolské kvalifikační práce jsou archiváliemi a pokud ano, jaký by měly mít skartační znak. Práce bakalářské, diplomové, rigorózní a disertační jsou školními díly podle Autorského zákona, vytvořenými pod vedením školy za účelem splnění studijních povinností. Mohou být také považovány za součást protokolu o státní závěrečné zkoušce studenta, který bezesporu archiválií je (6).

Odpověď na položenou otázku můžeme nalézt ve vzorových skartačních řádech, které jednotlivým dokumentům přiřazují odpovídající skartační znaky:

„A“ - dokumenty s trvalou dokumentární nebo informační hodnotou určené do trvalé úschovy v archivu,

„V“ - dokumenty, které jsou po uplynutí skartační lhůty znovu v rámci skartačního řízení posouzeny a zařazeny do kategorie dokumentů typu „A“ nebo „S“,

„S“ - dokumenty, u nichž z hlediska dokumentárního nebo informačního není po splnění jejich provozní a správní funkce nutná další úschova a mohou být v rámci skartačního řízení vyřazeny a skartovány.

Číselné označení uváděné za skartačním znakem udává skartační lhůtu v letech, po kterou musí být písemnost z provozních či správních důvodů v organizaci uložena. Po jejím uplynutí se dokument ve skartačním řízení navrhne k předání do příslušného archivu (skartační znak „A“) nebo ke zničení (skartační znak „S“) anebo k posouzení, má-li se písemnost předat do archivu nebo má-li se zničit (skartační znak „V“).

Podle výnosu ministerstva školství č. j. 11834/57 ze dne 16. 7. 1958 a č. j. 10324/60-L ze dne 1. 4. 1960 byly diplomové a jim podobné práce považovány za archiválie se skartačním znakem A10. Aktualizací ve výnosu Ministerstva školství ČSR č. j. 19 151/87-491 z roku 1987 byly nově diplomové práce uvedeny se skartačním znakem V20, disertační a habilitační práce se skartačním znakem A20. Podle předpisu z roku 1987 měl být kritériem výběru diplomových prací ve skartačním řízení přínos vědě a výzkumu. Tento výběr však nebyl na

školách prakticky prováděn. Seznamy platných předpisů v resortu školství, mládeže a tělovýchovy vydávané po roce 2005 tento předpis již přestaly uvádět.

Nahlédnutím do skartačních řádů jednotlivých škol (viz Tabulka 1 *Přehled skartačních znaků VŠKP podle skartačních řádů vybraných univerzit*) zjistíme, že praxe archivování vysokoškolských kvalifikačních prací se velmi liší. Společným prvkem, který můžeme vysledovat, je význam přikládaný pracím habilitačním, které jsou uváděny se skartačním znakem A5 nebo A10, výjimečně se skartační lhůtou 40 let (Vysoká škola ekonomická v Praze – VŠE). Tabulka dokládá různou váhu, kterou přikládají jednotlivé školy pracím vznikajícím na nižším stupni studia (bakalářské, diplomové). Např. Univerzita Palackého v Olomouci (UPOL) zvolila pro práce bakalářské a diplomové ve svém skartačním řádu znak S20, který tyto práce umožňuje vyřadit a skartovat po dvaceti letech bez provedení skartačního výběru. Mendlova zemědělská a lesnická univerzita v Brně (MZLU) a Univerzita Karlova (UK) zvolily skartační lhůtu pouhých pět let.

Tabulka 1 Přehled skartačních znaků VŠKP podle skartačních řádů vybraných univerzit (zdroj: vlastní zpracování)

	VŠE	UK	UTB	UPOL	MZLU
bakalářské	V20	A5	V20	S20	V5
diplomové	V20	A5	V20	S20	V5
disertační	V20	A5	A5	A10	A5
habilitační	A40	A5	A5	A10	A5

Do způsobu archivace se promítl i uplynulý přechod od sběru tištěných vysokoškolských kvalifikačních prací ke sběru elektronických verzí. Díky legislativním změnám nyní školy nemusejí uchovávat práce v tištěné podobě v kapacitně omezených knihovnách a archivech, ale mohou zpřístupňovat a dlouhodobě uchovávat pouze dokument digitální. Např. Vysoká škola ekonomická v Praze považuje za primární dokument elektronickou verzi a vybírá a archivuje práce v elektronické podobě. Listinné verze (pokud jsou vůbec požadovány k obhajobě) se vracejí autorovi práce nebo je pracoviště, na němž se práce obhajují, skartuje ve vlastní režii a v souladu se skartačním řádem.

Podle provedeného průzkumu v roce 2014 archivaci VŠKP má v ČR v předpisech podchyceno 16 veřejných vysokých škol (viz oddíl 3.4.3).

Při budování repozitářů eVŠKP musíme mít na paměti požadavek dlouhodobého uložení a zpřístupnění vysokoškolských kvalifikačních prací. V případě, že by se mělo jednat

o archivní kopii kvalifikační práce, jsou požadavky na dlouhodobé uložení o to přísnější než u běžné šedé literatury. Důvod pro dlouhodobé uložení dokládají i proběhlé kauzy plagiátorství, ve kterých bylo často u vysokoškolských kvalifikačních prací prokazováno zkopírování cizích textů až po mnoha letech po obhajobě.

2.1.3 Zveřejňování a sdělování eVŠKP veřejnosti

Před rokem 2006 bylo běžnou praxí půjčovat tištěné práce prezenčně v knihovně studentům dané vysoké školy na základě školní licence dle *Zákona 216/2006 Sb. (kterým se mění zákon č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů; zkráceně Autorský zákon)*. V případě, že škola podepisovala se studenty licenční smlouvu, bylo možné půjčovat práce s licencí i širší veřejnosti.

Autorský zákon s účinností od 1. 12. 2000 mimo jiné vymezil nové situace, kdy užitím díla nedochází k zásahu do autorského práva. Kvalifikačních prací se týká § 37 Knihovní licence, podle kterého „do práva autorského nezasahuje knihovna, archiv, muzeum, galerie, škola, vysoká škola a jiné nevýdělečné školské a vzdělávací zařízení,

d) půjčuje-li originály nebo rozmnoženiny obhájených diplomových, rigorózních, disertačních a habilitačních prací na místě samém, a to výhradně pro účely výzkumu nebo soukromého studia, pokud takové užití autor nevyloučil.“ (7 § 37) Toto užití však neumožňuje

1) zpřístupnit práce bakalářské,
2) zpřístupnění prací před obhajobou,
3) volné zpřístupnění eVŠKP na Internetu (sdělování díla veřejnosti podle § 18 Autorského zákona).

V rámci veřejných konzultací proto autor disertační práce podal Ministerstvu kultury návrh (8) k připravované novele Autorského zákona řešící problematiku zpřístupňování VŠKP. V rámci návrhu mj. doporučil

1) úpravu § 28 odst. 1) Autorského zákona – výjimky a omezení práva autorského by mělo být vhodné uplatnit i na základě dalších zákonů než jen podle Autorského zákona, příkladem je výjimka u eVŠKP v § 47b zákona o vysokých školách, viz níže;
2) úpravu § 37 odst. (1) d) – rozšíření výčtu VŠKP o práce bakalářské a povolení zpřístupňování prací před obhajobou.

V průběhu let 2012 – 2013 se autor disertační práce účastnil jednání zástupců MK ČR a zájmových organizací, během kterých byla novela zákona projednávána. Výše uvedené úpravy byly do návrhu novely přijaty, schvalování celého návrhu Autorského zákona však vzhledem k politické situaci na konci roku 2013 bylo odloženo a opětovně je projednáváno až v roce 2015.

Velmi významnou legislativní změnou byl zákon č. 552/2005 Sb. (5) (zákon o vysokých školách) zavádějící § 47b. Ten školám dává povinnost nevýdělečně zveřejňovat disertační, diplomové, bakalářské a rigorózní práce, u kterých proběhla obhajoba, včetně posudků oponentů a výsledku obhajoby. Tyto práce musí být též nejméně pět dní před obhajobou dostupné veřejnosti, je možné si z nich pořizovat výpisy, opisy nebo rozmnoženiny. Způsob, jakým jsou práce zveřejněny v databázi kvalifikačních prací, má stanovit vnitřní předpis vysoké školy. Autor odevzdáním své práce souhlasí s takovýmto zveřejněním bez ohledu na výsledek obhajoby. Zákon neupravuje zveřejňování prací habilitačních.

Novela vysokoškolského zákona v § 47b používá pojem *zveřejnit*, resp. *zveřejnit k nahlížení veřejnosti*. Není však jasné, co je těmito termíny míněno, neboť Autorský zákon zveřejňování k nahlížení veřejnosti nedefinuje. Definuje pouze zveřejnění díla v § 4 odst. (1):

„Prvním oprávněným veřejným přednesením, provedením, předvedením, vystavením, vydáním či jiným zpřístupněním veřejnosti je dílo zveřejněno.“ (7 § 4)

Pro potřeby zpřístupňování eVŠKP, tj. děl v elektronické podobě, by spíše odpovídal termín „sdělování veřejnosti“ definovaný v § 18:

„(1) Sdělováním díla veřejnosti se rozumí zpřístupňování díla v nehmotné podobě, živě nebo ze záznamu, po drátě nebo bezdrátově.
(2) Sdělováním díla veřejnosti podle odstavce 1 je také zpřístupňování díla veřejnosti způsobem, že kdokoli může mít k němu přístup na místě a v čase podle své vlastní volby zejména počítačovou nebo obdobnou sítí. (…)“ (7 § 18)

Podle § 47b Zákona o VŠ dále „odevzdáním práce autor souhlasí se zveřejněním své práce“, které „vysoká škola nevýdělečně zveřejňuje (…) prostřednictvím databáze kvalifikačních prací“ (5 § 47b). Podle Mgr. Věry Jurmanové Volemanové by příhodnější formulace,

v intencích Autorského zákona, měla být např. „autor díla uděluje oprávnění k výkonu práva dílo užít sdělováním díla veřejnosti“ (9).

Některé vysoké školy poukazovaly na možný rozpor s Autorským zákonem a s *Bernskou úmluvou o ochraně literárních a uměleckých děl*, požadovaly písemné uzavření smlouvy na užití díla mezi autorem a školou (udělení licence), případně práce nezpřístupňovaly v elektronické podobě bez omezení přístupu s tím, že není jasně definován pojem databáze kvalifikačních prací. Příkladem škol nezpřístupňujících práce online je Akademie múzických umění - AMU (student AMU nemá povinnost škole udělit licenci k užití díla), Slezská univerzita (zpřístupňuje veřejnosti práce pouze lokálně v knihovně) aj. viz *Průzkum zpřístupňování vysokoškolských kvalifikačních prací v roce 2014* v kapitole 3. Vzhledem k různorodé praxi ve zpřístupňování eVŠKP se autor přiklání k výkladu pojmu „zveřejňování“ v § 47b ve smyslu definice Open Access (viz oddíl 2.1.4), což odpovídá záměrům předkladatelů novely zákona o vysokých školách (resp. Výboru pro vědu, vzdělání, kulturu, mládež a tělovýchovu, který podal pozměňovací návrh zavádějící § 47b).

Vzhledem k uvedeným nejasnostem řešitelé projektu Národního úložiště šedé literatury (viz oddíl 2.3.2) nechali zpracovat právní expertizu (10) k digitálnímu zpracování šedé literatury, která se mj. zabývá analýzou právních vztahů plynoucích z elektronického zpracování, uchovávání a publikování kvalifikačních prací (autorských děl) včetně metadat zahrnutých v databázích univerzit. Podle kapitoly 3.2 zmiňované expertizy je sdělování VŠKP podle § 47b Zákona o vysokých školách zákonnou výjimkou, neboť je uzavřena licenční smlouva s autorem. „Tato licence má charakter tzv. implicitního dovolení, přičemž ust. novelou přidaného § 47b zákona č. 118/1998 Sb. ukládá vysokým školám zveřejňovat kvalifikační práce (zákonodárce zde ještě rozlišuje zveřejnění před obhajobou a po ní) – – z tohoto příkazu pak přímo ex lege vyplývá i dovolení užít příslušnou práci takto stanoveným způsobem.“ (10 str. 18)

V závěrečné části právní expertizy je doporučení pro vysoké školy uzavírat licenční smlouvy s autory eVŠKP konkludentně. Při konkludentním způsobu uzavření smlouvy je návrhem na uzavření licenční smlouvy vložení příslušného dokumentu do určené databáze, přijetím návrhu ze strany zadavatele je zkopírování tohoto dokumentu do vlastní databáze. Není potřeba uzavírat licenci písemně.

K výše zmíněnému pojmu databáze z § 47b uvedená analýza uvádí, že vysoké školy mají „povinnost (eVŠKP, pozn. autora) zveřejňovat, a to dle § 47b odst. 1 i prostřednictvím speciálních databází. Vysoké školy tedy mají v tomto případě, jak uvedeno shora, nejen implicitní zákonné dovolení tyto práce zveřejnit, ale též vést jejich databáze. Autorský zákon přitom s tímto pojmem přímo pracuje, když databázi v § 88 vymezuje jako ‚soubor nezávislých děl, údajů nebo jiných prvků, systematicky nebo metodicky uspořádaných a individuálně přístupných elektronickými nebo jinými prostředky, bez ohledu na formu jejich vyjádření. ' Vysoké školy se pak na základě zákonného příkazu (a jemu odpovídajícímu dovolení) v § 47b odst. 1 stávají pořizovateli databází.“ (10 str. 21)

Příslušná vysoká škola získává podle § 47b Zákona o VŠ pouze licenci k nevýdělečnému zveřejnění. Oprávněné subjekty (tj. kdokoli) si tedy na základě tohoto ustanovení mohou pořídit rozmnoženiny, které už však nesmí bez dalšího například dále šířit či zveřejňovat. Omezení dalšího šíření se týká např. dalšího zpřístupňování plných textů a metadat zprostředkovatelem, pokud by k tomu neměl oprávnění. V případě NUŠL je proto uzavírána s vysokými školami sublicence na vytěžování či zužitkování univerzitních repozitářů.

2.1.4 Definice Open Access

Vzhledem k prosazování otevřeného přístupu k výsledkům vědy a výzkumu (angl. „Open Access“), mezi které elektronické vysokoškolské kvalifikační práce můžeme počítat, je nutné upřesnit použití termínu otevřený přístup / Open Access v této disertační práci[2].

Česká terminologická databáze knihovnictví a informační vědy (11) v roce 2012 ještě správnou definici pojmu Open Access neuváděla, pod pojmem Open Access se nacházela definice související se svobodným přístupem k informacím. Na základě konzultace autora se správkyní databáze Jaroslavou Havlovou z Národní knihovny ČR, bylo v roce 2012 přislíbeno upravit termín tak, aby reflektoval současný význam termínů otevřený přístup k vědeckým informacím a svobodný přístup k informacím. Na konci roku 2014 již tato databáze nově obsahuje termín „otevřený přístup (k vědeckým informacím)“ (angl. ekv. „Open Access“), který je definován jako „Online přístup k odborným informacím, především k plným textům recenzovaných vědeckých článků, ale i k textům preprintů, konferenčních sborníků ad., bez

[2] Oddíl Definice Open Access vychází z autorovy seminární práce Metriky pro Open Access repozitáře (66).

poplatků a komukoli. Jeho hlavním cílem je dosáhnout větší (‚neomezené') možnosti šíření a zpřístupňování vědeckých poznatků pro odbornou, ale i laickou veřejnost v souladu s možnostmi, které poskytuje aktuální stav informačních technologií. Otevřený přístup se dělí dle uspořádání autorskoprávních vztahů na tzv. ‚volný otevřený přístup' a tzv. ‚bezplatný otevřený přístup'. Druhým typem dělení je dělení na dva základní publikační modely – dvě cesty naplnění otevřeného přístupu, a to na tzv. ‚zlatou cestu otevřeného přístupu' a tzv. ‚zelenou cestu otevřeného přístupu'. Definice otevřeného přístupu vychází z tzv. BBB-iniciativ (tj. *Budapešťská iniciativa*, *Prohlášení z Bethesdy* a *Berlínská deklarace*)." (11)

Výchozí definici otevřeného přístupu, aplikovatelnou i na otevřený přístup k eVŠKP v repozitářích, lze najít v *Budapešťské iniciativě*: „Literatura, která by měla být volně dostupná online, je ta, kterou vědci poskytují světu, aniž by za ni očekávali platbu. Primárně tato kategorie zahrnuje recenzované časopisecké články; patří sem ale i nerecenzované preprinty, které vědci mohou chtít nabídnout online pro připomínkování nebo jako upozornění kolegům na důležité výzkumné poznatky. Existuje mnoho stupňů a druhů širšího a snazšího přístupu k takové literatuře. Pojmem ‚otevřený přístup' k této literatuře myslíme její volnou dostupnost na veřejném internetu umožňující libovolnému uživateli číst, stahovat, kopírovat, distribuovat, tisknout, prohledávat nebo vytvářet odkazy na plné texty těchto článků, sklízet je pro potřeby indexace, předávat je jako data pro software, nebo používat je k jakýmkoliv jiným legálním účelům bez finančních, právních nebo technických omezení s výjimkou těch, která jsou neoddělitelnou součástí získání přístupu k internetu samotnému. Jediným omezením na reprodukci a distribuci a jediným uplatněním autorsko-právní ochrany (copyrightu) v této oblasti by mělo být poskytnout autorům kontrolu nad integritou jejich prací a právo na řádné uznání a uvedení autorství." (12)

Na *Budapešťskou iniciativu* navazují další, specifičtěji orientované. Např. v *Prohlášení z Bethesdy* (zaměřené na biomedicínu) lze nalézt požadavek na uložení kompletní verze práce i včetně všech částí v některém z online repozitářů: „Úplná verze práce a všech doplňkových materiálů, včetně souhlasu k využití, jsou ihned po prvotním publikování uloženy ve vhodném standardizovaném elektronickém formátu alespoň v jednom on-line repozitáři, který je podporován akademickou institucí, učenou společností, vládní agenturou nebo jinou dobře zavedenou organizací, která usiluje o poskytnutí otevřeného přístupu, neomezenou distribuci, interoperabilitu a dlouhodobou archivaci." (13)

Významným mezníkem v komunikaci vědeckých informací bylo zveřejnění *Berlínské deklarace o otevřeném přístupu ke znalostem v přírodních a humanitních vědách* (14), která má jako hlavní cíl podporu otevřeného přístupu k vědeckým poznatkům prostřednictvím Internetu. V definici *Berlínské deklarace* se vysloveně mluví o přístupu k plným textům.

V roce 2012 se institucionálním signatářem *Berlínské deklarace* v ČR stala Asociace knihoven vysokých škol ČR, což dokládá rostoucí význam Open Access publikování pro české vysokoškolské knihovny – podporu publikování akademických pracovníků, benchmarking úspěšnosti publikační činnosti vědců, univerzitních časopisů, vědců, vědeckých týmů. Na doporučení autora disertační práce se ve stejném roce připojila svým podpisem k *Berlínské deklaraci* i Vysoká škola ekonomická v Praze a potvrdila tak mj. svůj kladný postoj k otevřenému přístupu k vysokoškolským kvalifikačním pracím ve svém repozitáři (viz oddíl 2.3.2, část *Databáze kvalifikačních prací VŠE*).

Do prosince 2014 podepsaly *Berlínskou deklaraci* tyto české instituce (15):

- Grantová agentura ČR (2008)
- Akademie věd ČR (2008)
- Masarykova univerzita (2010)
- MAGNANIMITAS (2011)
- AKVŠ ČR (2012)
- Vysoká škola ekonomická v Praze (2012)
- Univerzita Karlova v Praze (2013)
- Výzkumný ústav komunikace v umění, o.p.s. (2013)
- Vysoké učení technické v Brně (2013)

Vzhledem k povinnosti škol zveřejňovat vysokoškolské kvalifikační práce dané novelou *zákona 111/1998 Sb., o vysokých školách a o změně a doplnění dalších zákonů* (5), která ukládá školám zveřejňovat eVŠKP (viz oddíl 2.1.3), bude nás v této práci zajímat reálná otevřenost přístupu do repozitářů eVŠKP jednotlivých veřejných vysokých škol v ČR ve smyslu výše uvedených definic Open Access, tj. v elektronické podobě, bezplatně, bez bariér, se zachováním autorství.

2.1.5 Definice plagiátorství

Tento oddíl obsahově vychází z autorova článku v časopise *ProInflow 2/2014* (19).

Základní definici plagiátorství najdeme v *ČSN ISO 5127:2003 Informace a dokumentace – – Slovník* (16), která stanovuje termíny pro usnadnění mezinárodní komunikace v oblasti informací a dokumentace a vybrané pojmy definuje. V této normě je plagiátem označeno „představení duševního díla jiného autora půjčeného nebo napodobeného v celku nebo z části, jako svého vlastního“ (16), tj. užití myšlenky bez uvedení jejího autora.

Autorský zákon (7) plagiátorství samotné nedefinuje, ale stanovuje autorské dílo v § 2 a související práva osobnostní v § 11 a majetková v § 12 až § 27, která upravují, jak je možné s dílem nakládat.

„Výjimky a omezení práva autorského lze uplatnit pouze ve zvláštních případech stanovených v tomto zákoně a pouze tehdy, pokud takové užití díla není v rozporu s běžným způsobem užití díla a ani jím nejsou nepřiměřeně dotčeny oprávněné zájmy autora.“ (7 § 29) Těmito výjimkami jsou především Volné užití a zákonné licence v § 30 až § 39. Pro vysoké školy je nejvýznamnější Citace jakožto volné užití podle § 31, podle kterého „Do práva autorského nezasahuje ten, kdo

a) užije v odůvodněné míře výňatky ze zveřejněných děl jiných autorů ve svém díle,
b) užije výňatky z díla nebo drobná celá díla pro účely kritiky nebo recenze vztahující se k takovému dílu, vědecké či odborné tvorby a takové užití bude v souladu s poctivými zvyklostmi a v rozsahu vyžadovaném konkrétním účelem,
c) užije dílo při vyučování pro ilustrační účel nebo při vědeckém výzkumu, jejichž účelem není dosažení přímého nebo nepřímého hospodářského nebo obchodního prospěchu, a nepřesáhne rozsah odpovídající sledovanému účelu;

vždy je však nutno uvést, je-li to možné, jméno autora, nejde-li o dílo anonymní, nebo jméno osoby, pod jejímž jménem se dílo uvádí na veřejnost, a dále název díla a pramen.“ (7 § 31)

Jakékoliv jiné užití než v Autorském zákoně uvedené je tedy porušením Autorského zákona. U citací se jedná především o tyto přestupky:

a) použití většího rozsahu cizího textu než je míra odůvodněná,

b) použití cizího celého díla při výuce nebo vědeckém výzkumu za účelem hospodářského nebo obchodního prospěchu (překročení zákonné licence),
c) neuvedení jména autora, názvu díla či pramene, ať již úmyslné či neúmyslné (nevědomé plagiátorství).

Pokud dojde k odhalení projevů plagiátorství na vysoké škole před ukončením studia, podle vnitřních předpisů školy může být toto jednání podstoupeno např. disciplinární komisi. Na Vysoké škole ekonomické v Praze je postih za plagiátorství řešen v *Disciplinárním řádu pro studenty fakult Vysoké školy ekonomické v Praze*, který při zaviněném „porušení povinností stanovených právními předpisy nebo vnitřními předpisy VŠE a fakulty“ (17) umožňuje uložit jako sankci napomenutí, podmíněné vyloučení ze studia nebo vyloučení ze studia.

V případě, kdy student již úspěšně obhájil závěrečnou kvalifikační práci, ve které se dopustil plagiátorství a získal tak neoprávněně vysokoškolský titul, není již možné získaný titul odebrat nebo uplatnit jinou sankci podle vnitřních předpisů školy. Může se však stále občanskoprávní žalobou bránit autor původního, zneužitého textu. Podle § 40 Autorského zákona se může domáhat zejména určení svého autorství, zákazu ohrožení svého práva, odstranění následků (tj. například stažení zveřejněné plagiátorské práce nebo uvedení pravého autorství textu) a poskytnutí přiměřeného zadostiučinění za způsobenou nemajetkovou újmu (zejména omluvu a příp. finanční náhradu určenou soudem, pokud by se jiné zadostiučinění nejevilo postačujícím).

Na základě celosvětového dotazníkového šetření mezi 900 středoškolskými a vysokoškolskými vyučujícími (18) bylo identifikováno deset různých projevů plagiátorství:

1) klonování – vydávání cizí práce, slovo od slova, za vlastní,[3]
2) CTRL-C – vydávání cizí práce za vlastní, s minimálním množstvím úprav,
3) najít/nahradit – změna klíčových slov a frází bez změny podstaty textu,
4) remixování – parafrázování, spojení více zdrojů do jednoho,
5) recyklování – využití předchozích textů autora, bez autocitace,
6) hybridní – mixování velmi dobře citovaných zdrojů s necitovanými,

[3] Jako tento způsob plagiátorství lze chápat i doslovné přeložení zdroje bez řádné citace. Překládání z cizího jazyka bez citace je u českých textů běžnější než u textů anglických, úkolem testu v kapitole 8 bude mj. ověřit schopnosti antiplagiátorských nástrojů odhalit i překlady.

7) míchání zdrojů – kombinace více necitovaných zdrojů do jednoho textu,
8) chyba 404 – citace neexistujících zdrojů nebo špatné informace o zdroji,
9) agregace – korektní citování cizích zdrojů, téměř bez vlastního osobního přínosu autorem,
10) re-tweet – korektní citování, ale za využití originálního textu/struktury bez podstatnějších úprav.

Mezi nejčastější prohřešky patří podle průzkumu metody klonování, CTRL-C a míchání zdrojů, kdy nedochází k podstatným úpravám originálního textu. Jednotlivé softwarové nástroje analyzované v kapitole 8 pomáhají odhalit právě takovéto projevy plagiátorství díky nalezení duplicitních pasáží mezi analyzovaným textem a dalšími indexovanými dokumenty. Zda došlo k plagiátorství, musí rozhodnout recenzent posouzením nalezených duplicit. Aplikace samotné nedokáží analýzou psaného textu posoudit, zda se jedná o korektní citaci zdroje či např. o obecnou frázi, u které citace zdroje nutná není. Označení aplikací jakožto „systémů na vyhledávání plagiátů“ je tedy terminologickou nepřesností, v práci proto většinou hovoříme o antiplagiátorských systémech či systémech na vyhledávání duplicit.

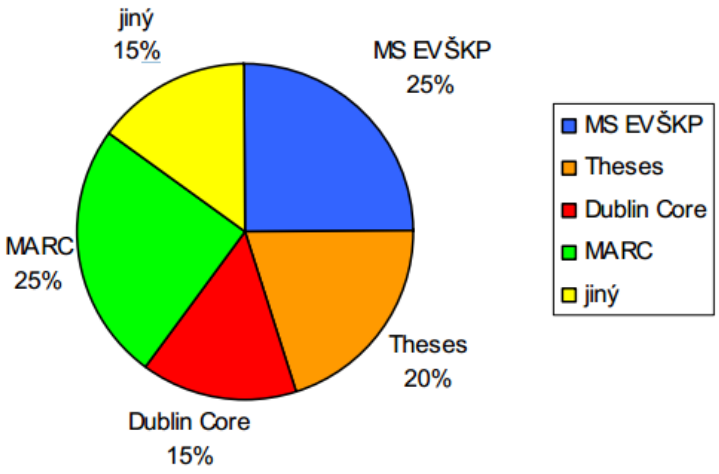
2.2 Zpřístupňování eVŠKP do roku 2014

Posledního dotazníkového šetření aktuálního stavu problematiky VŠKP (21), který provedla na podzim 2009 Odborná komise pro otázky elektronického zpřístupňování vysokoškolských kvalifikačních prací AKVŠ ČR (zkráceně Komise eVŠKP), se zúčastnilo pouze 16 škol oproti 26 v průzkumu v roce 2007.

Některé školy v roce 2009 práce vybíraly i nadále pouze v tištěné podobě a zpřístupňovaly je v rámci jednotlivých fakult. Některé se rozhodly kvalifikační práce zpřístupňovat v rámci školy nebo knihovny s odkazem na zákonnou licenci danou Autorským zákonem. U těchto škol byl zájemce většinou nucen stát se registrovaným čtenářem příslušné vysokoškolské knihovny, pokud chtěl k pracím získat přístup. Třetí kategorie škol práce zpřístupnila volně na Internetu, a to buď s odkazem na souhlas autora daný dle vysokoškolského zákona odevzdáním práce, nebo pro tyto účely i nadále získávaly licence od studentů (v průzkumu roku 2009 jeden respondent, ostatní školy zpřístupňující práce nebyly v dotazníku zahrnuty).

Z průzkumu vyplynulo, že metadatový formát pro popis VŠKP záleží na umístění záznamů. Na 41 % institucí zpřístupňovalo záznamy prostřednictvím knihovního katalogu. Více než

polovina institucí používala specializovaný software, nejčastěji DSpace (31 %). Z uvedených důvodů byl také formát použitý pro popis metadat odlišný, nejčastěji se jednalo o metadatový standard EVSKP-MS navržený Komisí eVŠKP nebo knihovní formát MARC (shodně po 25 %), následovaný proprietárním formátem pro Theses.cz (20 %) a Dublin Core (15 %), viz Obrázek 1 (popis metadatových formátů viz kapitola 4 *Mapování metadat eVŠKP*).



Obrázek 1 Dostupné formáty popisu eVŠKP (20)

Pro předávání metadat do národního registru VŠKP Theses.cz se využíval v 60 % proprietární formát Theses.cz a ve 40 % formát EVSKP-MS. Pouze 15 % škol sdílelo metadata podle protokolu OAI-PMH, nejčastěji byla předávána nahráním dávkově (47 %) nebo ručně (38 %).

Z průzkumu dále vyplynulo, že na Univerzitě Karlově zpřístupňování VŠKP spadalo do kompetence jednotlivých fakult a knihoven, většinou zde neexistovalo workflow odevzdávání eVŠKP.

Činnost Komise eVŠKP byla ukončena po naplnění jejích cílů 30. dubna 2010. Komise doporučila následné řešení problematiky evidence, dlouhodobého uchování a podpory zpřístupňování výsledků vědy a výzkumu z produkce vysokých škol na nové platformě, v rámci Odborné komise pro digitální repozitáře vysokých škol. V záměru předloženém výkonnému výboru AKVŠ ČR bylo autorem disertace, jakožto předsedou končící Komise eVŠKP, navrženo následující zaměření komise nové:

1) Sledování aktuálních vývojových trendů v zahraničí a v ČR v oblasti budování, provozování a dalšího rozvíjení repozitářů publikované i nepublikované literatury

z produkce vysokých škol a souvisejících informačních zdrojů vědy, výzkumu a vzdělávání a také návazných informačních služeb.

2) Podpora zpřístupňování produkce vědy, výzkumu a vzdělávání na vysokých školách prostřednictvím repozitářů, s důrazem na otevřený přístup (Open Access).
3) Sledování standardů a technologií souvisejících s provozem repozitářů pro vědu, výzkum a vzdělávání na vysokých školách (metadata, datové formáty, OAI-PMH, trvalé identifikátory aj.), doporučování zavádění standardů a technologií do praxe ČR.
4) Sledování problematiky dlouhodobého uchovávání a zpřístupňování informační produkce z oblasti vědy, výzkumu a vzdělávání na vysokých školách.

Založení nové komise výkonným výborem AKVŠ ČR nebylo schváleno pro údajný nezájem škol. V důsledku toho došlo během 3. setkání českých uživatelů systému DSpace ke vzniku nezávislé pracovní skupiny pro podporu Open Access. Tato skupina zástupců knihoven a univerzit od roku 2010 pravidelně pořádá setkání a koordinuje přípravu akce *Týden Open Access* na jednotlivých univerzitách členů. Zároveň byl vytvořen informační portál Open Access v ČR http://www.openaccess.cz, který spravuje autor disertace.

Kvůli vzrůstající problematice plagiátorství se do centralizovaného rozvojového projektu MŠMT ČR na zpřístupňování eVŠKP (viz Theses.cz, oddíl 2.3.2) zapojily další školy a stávající provedly mnoho změn ve svých digitálních repozitářích. Po ukončení činnosti Komise eVŠKP však již nebyl realizován srovnatelný průzkum zaměřený na problematiku sběru a zpřístupňování eVŠKP v ČR.

Na průzkumy realizované Komisí eVŠKP v letech 2006, 2007 a 2009 dílčím způsobem navázaly především autorky závěrečných prací na Masarykově univerzitě a na Univerzitě Karlově. Analýza kvalifikačních prací ke zpřístupňování eVŠKP je uvedena níže.

V bakalářské práci (21) autorka Tereza Balabánová popisovala technické aspekty zpřístupňování eVŠKP, popis existujících systémů a vznikající systém Masarykovy univerzity. Popis technických prostředků zůstal na obecné úrovni (XML, PDF, Dublin Core), v části popisující zahraniční a české systémy se autorka zaměřila pouze na vybrané příklady. Stěžejní část práce je věnována popisu systému a procesů zpracování VŠKP na Masarykově univerzitě. Autorka se často odvolávala na výsledky Komise eVŠKP. Vzhledem k datu publikování (rok 2006) a omezenému okruhu sledovaných případů výsledky studie pro nás již nejsou relevantní.

Autorka Šárka Absolonová ve své bakalářské práci (22) poukázala na problém různorodosti řešení zpřístupňování informací i samotných prací na lokálních úrovních. Zdůraznila potřebu sjednocení norem, stanovení metodických pokynů, standardů a pravidel (o což usilovala i, podle autorky, Komise eVŠKP). V souvislosti se zpřístupňováním eVŠKP veřejnosti upozornila na riziko plagiátorství, jako příklad řešení rizika je uváděn v práci popisovaný systém na odhalování plagiátů Masarykovy univerzity. V práci se zabývala také legislativními povinnostmi a omezeními, v průzkumu např. využíváním licenčních smluv, doporučenými formáty, popisnými metadaty, workflow a praxí na jednotlivých školách. Ve své práci se mj. při popisu repozitářů na vybraných školách často odvolávala na výstupy Komise eVŠKP a jejích členů. Vzhledem k datu publikování (rok 2008) a k využití již dříve publikovaných výsledků Komise eVŠKP výsledky studie opět nemůžeme brát za aktuálně platné.

Významnější prací obhajovanou na Univerzitě Karlově je diplomová práce Růženy Zlatohlávkové *Digitální repozitáře na vysokých školách v České republice* (23), která v kapitole 5 představuje podrobnou analýzu vybraných digitálních repozitářů vysokých škol v ČR, včetně systémů DSpace. Drobnými nedostatky práce, kritizovanými i v oponentním posudku vedoucí práce, je navržené dělení repozitářů podle využívaného software a výběr repozitářů (např. za Univerzitu Karlovu je zmíněn pouze Repozitář Digitool, ale ne již novější Repozitář závěrečných prací viz oddíl 2.3.2).

Irena Baranyová, která byla členkou Komise eVŠKP, se ve své diplomové práci (24) zabývala problematikou vysokoškolských kvalifikačních prací v elektronické podobě a důvody jejich zpřístupňování. Samostatné kapitoly popisují situaci na Slovensku (dnes již neaktuální systém ETD.SK), ve Velká Británii (systém EThOS) a v České republice (Digitální repozitář UK, Theses.cz). V poslední kapitole jsou jednotlivé systémy porovnány. Pro průzkum byla použita převážně analýza volně dostupných informací, v analýze situace v ČR autorka vycházela mj. z dokumentů Komise eVŠKP. Uvedené poznatky o repozitáři Univerzity Karlovy je zapotřebí aktualizovat (viz oddíl 2.3.2), neboť proces zpřístupňování na univerzitě v posledních pěti letech prošel zásadní změnou.

Pro téma této disertační práce je zajímavá diplomová práce Jitky Bugajevové (25). V teoretické části autorka uvádí související právní předpisy, na příkladu vybraných vysokých škol popsala situaci v oblasti zpřístupňování eVŠKP (s odkazem na činnost Komise eVŠKP a projekt Národního úložiště šedé literatury) a ve srovnání se stavem na Slovensku porovnala

systémy MUNI (Theses.cz a Odevzdej.cz) na kontrolu kvalifikačních prací proti plagiátům. Významná je též praktická část věnovaná kvalitativnímu průzkumu dostupnosti kvalifikačních prací na českých vysokých školách. Výzkum byl zaměřen na praktické možnosti veřejnosti získat plné texty tištěných či elektronických verzí závěrečných prací na vybraných veřejných, státních a soukromých vysokých školách. Autorka zvolila při sběru dat účelové vzorkování – extrémní variantu (26), která ze své podstaty nemusí reprezentovat stav na českých univerzitách, ale pouze extrémní situace. „Zařadila jsem do svého výzkumu naše dvě největší školy – Univerzitu Karlovu v Praze a Masarykovu univerzitu. (…) Do svého výzkumu jsem se snažila zařadit různé typy škol z hlediska jejich zaměření např. technické, umělecké apod. Bohužel některé veřejné vysoké školy zájem o výzkum neprojevily (např. České vysoké učení technické v Praze, Technická univerzita v Liberci a Vysoké učení technické v Brně). Vybírala jsem i mezi soukromými školami. Zde jsem se snažila získat pro svůj výzkum ty vysoké školy, které mají nejvíce studentů, opět z důvodu předpokladu nejpropracovanějšího systému nakládání s VŠKP. (…) Výzkumu se zúčastnilo celkem 13 vysokých škol (4 soukromé, 1 státní a 8 veřejných).“ (26 str. 67)

Autorka použila strukturovaný rozhovor s otevřenými otázkami. Analyzované okruhy a hlavní poznatky byly:

1) Koordinace problematiky zpřístupňování VŠKP
   a) studenti všech škol odevzdávají elektronické verze VŠKP
   b) workflow VŠKP je většinou koordinováno centrálně, za zpřístupnění zodpovídají knihovny
2) Retrospektivní digitalizace VŠKP
   a) retrospektivní digitalizace VŠKP je problematická z autorskoprávních důvodů, dále z finančních, personálních a kvůli malé poptávce
3) Licenční smlouvy
   a) z respondentů pouze AMU uzavírá licenční smlouvy
   b) student může na školách požádat o nezpřístupnění své práce
4) Typy VŠKP
   a) netextové typy prací se vyskytují na několika veřejných vysokých školách
   b) zúčastněné školy evidují min. bakalářské a magisterské práce (autorka se domnívá, že soukromé vysoké školy nemají disertační práce z důvodu své krátké působnosti, důvodem bude spíše absence akreditace doktorského studia)
5) Tištěné VŠKP
   a) tištěné práce jsou většinou zpřístupňovány v knihovně prezenčně, dokud nejsou vyřazeny

6) eVŠKP
   a) povinnost zpřístupňovat elektronickou verzi VŠKP je u všech respondentů řešena na celouniverzitní úrovni
   b) soukromé školy nejsou ochotny zveřejňovat plné texty VŠKP v elektronické podobě, plné texty zveřejňují jen některé veřejné vysoké školy, přístup je někdy podmíněn registrací
7) Meziknihovní výpůjční služba
   a) meziknihovní výpůjční služba pro VŠKP není poskytována
   b) většina knihoven je ochotna na požádání zaslat okopírovaný či naskenovaný obsah a bibliografii, eventuálně i několik stran textu
8) Kontrola kvalifikačních prací
   a) většina škol využívá systém Theses.cz, jedna soukromá škola používá systém Ephorus
   b) některé školy poskytují metadata do Národního úložiště šedé literatury

Vzhledem k neochotě soukromých vysokých škol zveřejňovat eVŠKP se jeví vhodnou volba veřejných vysokých škol jakožto cílové skupiny připravovaného průzkumu, v rámci kterého je žádoucí ověřit hypotézu, že povinnost zpřístupňovat elektronické verze je na všech školách řešena na celouniverzitní úrovni (tj. zda ve studii Jitky Bugajevové nedošlo ke zkreslení účelovým výběrem vzorku – 8 veřejných vysokých škol), a zjistit, jaké předpisy na školách problematiku zpřístupňování řeší.

Výše uvedená analýza pramenů ukazuje význam již publikovaných dotazníkových šetření a dalších výstupů Komise eVŠKP. Na základě analýzy dostupných zdrojů však lze usuzovat, že po ukončení činnosti Komise eVŠKP již nedošlo ke komplexnějšímu průzkumu stavu zpřístupňování eVŠKP v ČR a vzhledem k zájmu odborné veřejnosti o problematiku je vhodné naši znalost aktualizovat.

2.3 Repozitáře VŠKP

Současná praxe a trendy ve vyhledávacích rozhraních eVŠKP byly zjišťovány na základě studia významných českých a zahraničních repozitářů eVŠKP. V následující části je uvedena charakteristika vybraných repozitářů eVŠKP, které jsou významné svým obsahem, které byly vzorem pro práci v Komisi eVŠKP při návrhu workflow a metadat nebo které jinak přispěly k výzkumu v této práci.

Uvedený přehled bude využit v kapitole 7 jako podkladová analýza pro následný návrh modelové aplikace pro indexování a vyhledávání eVŠKP, jednotlivé repozitáře jsou proto popsány s důrazem na uživatelské rozhraní pro vyhledávání eVŠKP. Jednou ze sledovaných

vlastností bude využití fasetového vyhledávání s klasifikací nalezených záznamů do relevantních subkategorií, které uživatel může vybírat v libovolném pořadí a zužovat tak seznam nalezených výsledků. Dle provedených průzkumů vybraných rozhraní nové generace OPAC a discovery služeb, fasetové vyhledávání uživatelé preferují, považují za intuitivní, nápomocné, užitečné pro zpřesňování dotazů a objevování obsahu katalogů. (27)

Při výběru a analýze významných zahraničních repozitářů autor vycházel mj. z prezentací a neformálních diskusí na International Symposium on Electronic Theses and Dissertations (zkráceně ETD), pořádaných organizací Networked Digital Library of Theses and Dissertations. Konferencí ETD se autor účastnil v letech 2007 – 2010 a v roce 2014.
Na konferenci v Pittsburghu v roce 2009 prezentoval poster (28) na téma aktuální stav zpřístupňování českých VŠKP a šedé literatury v projektu NUŠL.

2.3.1 Zahraniční repozitáře

Mezi významné zahraniční projekty a s nimi související portály eVŠKP patří mezinárodní NDLTD Union Archive s rozhraním VTLS Theses Search / Chamo Discovery, ProQuest Dissertations & Theses, britský systém EThOS a evropský projekt DART-Europe E-theses Portal.

NDLTD Union Archive

Koncept elektronických kvalifikačních prací byl poprvé diskutován v USA v roce 1987 na setkání v Michiganu. Rozpracován byl na začátku 90. let minulého století, především na Virginia Polytechnic Institute and State University (zkráceně Virginia Tech) pod vedením profesora Edwarda A. Foxe, který se stal spolupředsedou pracovní skupiny pro diplomové práce, technické zprávy a disertační práce.

Účastníci workshopu SURA v roce 1994 zvolili Adobe Portable Document Format (zkráceně PDF) a Standard Generalized Markup Language (zkráceně SGML) pro reprezentaci a archivaci prací. Na základě proběhlých odborných diskusí a workshopů vznikl v roce 1996 na Virginia Tech univerzitě software ETD db, zajišťující kompletní řešení pro odevzdávání, zpracování, archivaci a zpřístupnění kvalifikačních prací. Program byl uvolněn pro bezplatné užívání na mezinárodní úrovni.

Virginia Tech koordinovala vývoj a implementaci systému distribuované digitální knihovny, která by agregovala data vybíraná jednotlivými spolupracujícími institucemi. Systém umožňoval prohlížení a vyhledávání podle instituce, data vytvoření, autora, názvu, téma práce a plného textu práce. Plné texty kvalifikačních prací byly zpřístupněny celosvětově s možností stažení, uložení a tisku. Provoz vyhledávače se potýkal s běžnými problémy metavyhledávačů – aplikace nepodporovala deduplikace záznamů, rychlost vyhledávání odpovídala rychlosti nejpomalejšího lokálního repozitáře, při změně lokálního rozhraní muselo dojít k přenastavení federativního vyhledávače. (29)

V roce 1996 byla založena skupina National Digital Library of Theses and Dissertations. Po rozšíření své působnosti za hranice USA byla přejmenována na Networked Digital Library of Theses and Dissertations (zkráceně NDLTD). V roce 2003 se z ní stala nezisková charitativní organizace, která v současnosti sdružuje již stovky univerzit z celého světa a partnerských organizací.

Nejvýznamnějším dokumentem NDLTD je standard *ETD-MS: an Interoperability Metadata Standard for Electronic Theses and Dissertations* (30), který popisuje vysokoškolské kvalifikační práce za využití 13 Dublin Core prvků a vlastního prvku thesis.degree, obsahujícího podprvky name, level, discipline a grantor. Současná verze 1.1 standardu z 19. 8. 2010 obsahuje definici prvků ve formátu XML Schema a příklady zápisu metadat včetně mapování prvků ETD-MS do MARC 21 a MARCXML. Standard ETD-MS ve verzi 1.0 sloužil jako východisko pro návrh dalších zahraničních standardů včetně českého EVSKP-MS (viz kapitola 4 *Mapování metadat eVŠKP*).

Díky standardizaci přenosu metadat ve formátu XML přes protokol HTTP – specifikaci *Open Archives Initiative Protocol for Metadata Harvesting* (31) – byl umožněn vznik nového typu vyhledávacích služeb, které umožnily dávkové nahrávání metadat z jednotlivých repozitářů do centrálního úložiště a následné vyhledávání nad tímto úložištěm. Protokolu OAI-PMH bylo v NDLTD využito pro univerzálnost, jednoduchost a rozšiřitelnost o další metadatové standardy. Pro potřeby evidence metadat eVŠKP NDLTD byl zprovozněn repozitář NDLTD Union Archive (http://union.ndltd.org/) sdružující metadata ze zapojených repozitářů. Pro kódování znaků bylo použito znakové sady unicode - UTF-8 podporující i neanglické znaky. Union Archive slouží nejen jako OAI-PMH klient agregující metadata z repozitářů, ale i jako OAI-PMH server poskytující metadata dalším zájemcům. Krátce po vytvoření v únoru 2002

archiv obsahoval 5 346 prací z 9 kolekcí (z toho pouze 4 podporovaly standard ETD-MS) (29), v lednu 2015 obsahuje již téměř 4 miliony záznamů.

Díky dodržování standardu ETD-MS a zpřístupnění metadat z NDTLD Union Archive protokolem OAI-PMH je uživatelům k dispozici na výběr více vyhledávacích systémů, kromě univerzálních vyhledávačů zaměřených na vědecké publikace typu Google Scholar, Microsoft Academic Search a Scirus[4] společnosti Elsevier je to především vyhledávací rozhraní VTLS Theses Search se službou Chamo discovery.

VTLS Theses Search a Chamo Discovery

Původní vyhledávací rozhraní nad OPAC NDLTD bylo vytvořeno ve spolupráci Virginia Tech a VTLS Inc. V roce 2000 umožňovalo vyhledávání metadatových záznamů podle autora, názvu práce, věcného popisu, člena zkušební komise (vedoucího práce/oponenta), čísla knihovny anebo podle názvu časopisu, příp. později podle instituce (dostupná vyhledávací pole byla v čase měněna).

Nová verze vyhledávání založená na systému VTLS Visualizer byla prezentována mj. na konferenci ETD2009 v Pittsburghu (32) a následně autorem disertace v ČR představena české odborné veřejnosti v rámci 4. ročníku semináře Systémy pro zpřístupňování VŠKP (33). Discovery systém VTLS Visualizer umožňoval omezení výsledkové množiny pomocí faset (jazyk dokumentu, formát, rok obhajoby, kontinent a země aj.) s grafickým zvýrazněním rozdělení četností u jednotlivých faset.

VTLS Visualizer byl posléze nahrazen discovery systémem Chamo Discovery stejné společnosti, který je používán dodnes. Chamo Discovery podporuje tvorbu uživatelských seznamů, použití štítků a hodnocení jednotek, uživatelské komentáře, hodnocení a sdílení odkazů prostřednictvím sociálních sítí. Aktuální implementace pro NDLTD umožňuje třídění podle relevance, data přidání, autora, názvu, roku vytvoření či publikování anebo podle interního čísla knihovny. Výsledkový seznam je možné zúžit pomocí faset pro zdroj metadat, jazyk, formát, rok publikování, univerzity nebo věcný popis. Detailní záznam kvalifikační práce obsahuje název práce, jméno autora, odkaz na zdrojový repozitář a abstrakt práce.

[4] Provoz služby Scirus společnost Elsevier ukončila v březnu 2014.

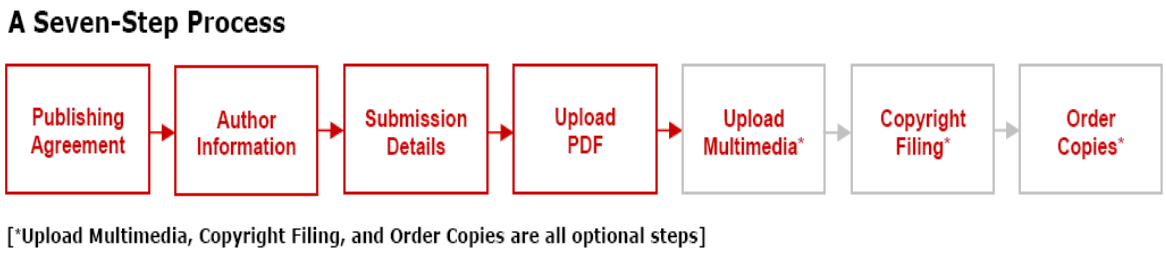
ProQuest Dissertations & Theses

V roce 1938 byla Eugenem Powerem založena společnost University Microfilm Inc. (zkráceně UMI) za účelem uchovávání knihovních sbírek na mikrofilmech. V roce 1939 se společnost začala věnovat publikování disertací. V roce 1951 UMI získala kontrakt od ALR na mikrofilmování pro databázi Dissertations Abstracts, které později začala vydávat na CD-ROM a digitalizovat. V roce 1998 Kongresová knihovna stanovila UMI jako svůj oficiální repozitář disertačních prací. (34, 35)

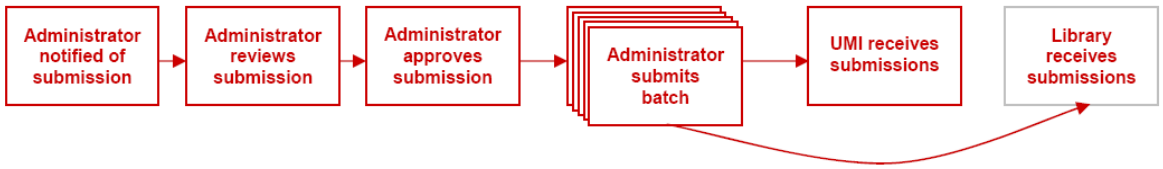
Nástupcem databáze ProQuest Dissertations Abstracts International se stala online databáze ProQuest Dissertations & Theses (zkráceně PQDT) s fulltextovým vyhledávání uložených prací. V současné době databáze obsahuje (35) na 3 miliony disertačních a dalších kvalifikačních prací od roku 1973 do současnosti, z toho 1 milion prací dostupných ke stažení ve formátu PDF. Roční přírůstek dokumentů s plným textem činí okolo 80 000. Metadata prací v PQDT jsou vytvářena a kontrolována editory ProQuest.

Vyhledávací rozhraní PQDT umožňuje fulltextové vyhledávání s možností filtrace pouze dokumentů s PDF aj. faset, řazením podle relevance nebo data publikování, vyhledávání podobných dokumentů, tvorbu uživatelských profilů s vlastními seznamy a sdílení prostřednictvím e-mailu a na sociálních sítích. Součástí kvalifikačních prací mohou být doplňující materiály s max. velikostí 10 GB, např. audio, video, data a tabulky, prezentace aj.

Kvalifikační práce v PQDT jsou dostupné předplatitelům např. v jednotlivých oborově zaměřených kolekcích nebo jako součást dalších produktů ProQuest. Instituce mohou využít online rozhraní ETD Administrator pro sběr a řízený proces odevzdání kvalifikačních prací školy (workflow viz Obrázek 2 a Obrázek 3), kdy začlenění práce do kolekce PQDT je bez poplatků. Autoři mohou zveřejnění práce podmínit časovým embargem nebo naopak práci zpřístupnit v režimu Open Access v rámci volně přístupné databáze PQDT Open (http://pqdtopen.proquest.com/).



Obrázek 2 Workflow pro vložení práce studentem v PQDT ETD Administrator (114)



Obrázek 3 Workflow schvalování práce administrátorem v PQDT ETD Administrator (114)

V rámci průzkumu uživatelů databáze PQDT a jejich chování (36), prezentovaného na semináři Komise eVŠKP v roce 2009, byly mimo jiné identifikovány následující poznatky:

- „Přes 3/4 (76 %) uživatelů PQDT jsou studenti. Téměř polovina (45 %) jsou postgraduální a třetina (31 %) pregraduální. Postdoktorandi a akademičtí pracovníci (10 %), vědci (6 %) a knihovníci (7 %) jsou občasnými uživateli.
- Téměř polovina (46 %) uživatelů PQDT buď pracují na své diplomové práci (13 %) nebo studují na doktorát (33 %). Většina uživatelů tedy nepracuje na diplomové nebo dizertační práci: jsou již absolventy PhD nebo jsou pregraduálními studenty.
- Téměř polovina (47 %) uživatelů PQDT vyhledávají přímo dizertace. Většina pregraduálních studentů (71 %) pravděpodobněji než ostatní vyhledávali obecněji a jejich výsledky zahrnovaly i dizertace.
- Většina (55 %) uživatelů PQDT by dalo přednost zpřístupnění podkladových dat k diplomovým a dizertačním pracím.“ (36)

Lze předpokládat, že výše uvedené výsledky průzkumu budou vypovídající i o chování uživatelů dalších repozitářů eVŠKP.

EThOS

Ve Velké Británii je národním systémem pro online zpřístupnění disertačních prací Electronic Theses Online System (zkráceně EThOS) (37). EThOS byl vyvinut jakožto nástupce služby Britské knihovny British Thesis, založené na digitalizaci prací na mikrofilmy a následném půjčování mikrofilmů. Cílem systému EThOS je maximalizovat dostupnost a viditelnost disertačních prací pomocí jednoho vyhledávacího rozhraní.

Měsíčně je evidováno na 75 000 návštěv webu a zpřístupněno na 30 000 plných textů (37). V lednu 2015 systém EThOS indexoval přes 380 000 prací, z toho přibližně ¼ byla dostupná k okamžitému stažení v elektronické verzi.

Koncepce provozu systému je popsána v souboru dokumentů *EThOS Toolkit* (38), zahrnuje popis správy systému, digitalizace, metadat a související právní problematiky.

Systém shromažďujeme metadata z více jak 130 zapojených univerzit protokolem OAI-PMH v proprietárním metadatovém formátu EThOS UKETD_DC (viz kapitola 4 *Mapování metadat eVŠKP*), který se skládá z metadatových prvků Dublin Core a rozšíření o prvky ze jmenného prostoru uketd_dc (39). Mezi základní prvky, které musí zapojený repozitář publikovat, patří autor, název práce, název instituce přidělující titul, datum zkoušky a typ studijního programu (EThOS indexuje pouze práce z doktorského studia). Alternativně lze metadata zaslat na CD-ROM nebo DVD.

Vyhledávací rozhraní nabízí základní a pokročilé vyhledávání. V základním vyhledávání se prohledávají metadata a dostupné plné texty na všechna zadaná klíčová slova. Použitá syntaxe podporuje vyhledání frází použitím uvozovek. Výsledky vyhledávání je možné omezit pouze na práce s dostupným plným textem.

Pokročilé rozhraní nabízí vyhledávání s upřesněním prohledávaného metadatového pole s výběrem pole z rozbalovacího seznamu. Termíny zadané v jednom řádku jsou automaticky použity jako vyhledávaná fráze v daném poli. Jednotlivé řádky v pokročilém vyhledání lze spojit pomocí booleovských operátorů AND, OR a AND NOT, nové řádky lze libovolně přidávat tlačítkem Add a Row. Seznam polí obsahuje na 20 různých položek, kromě běžných polí typu abstrakt, název disertace, instituce přidělující titul, jméno či příjmení autora, vedoucí práce a rok či přesné datum obhajoby je možné vyhledat i podle Etheses ID, perzistentního

identifikátoru EThOS, předmětových hesel Kongresové knihovny (LCSH), signatury, jazyka práce aj. polí.

Zobrazený metadatový záznam práce obsahuje název práce, jméno autora, instituci přidělující titul, instituci zpřístupňující práci, datum přidělení titulu, dostupnost plného textu, perzistentní identifikátor EThOS, jméno vedoucího práce, přispívající osoby, typ studijního programu a titul, abstrakt a klíčová slova. Nalezený záznam je možné sdílet nejen přes e-mail, ale i na sociálních sítích Facebook, Twitter, LinkedIn, CiteULike aj.

U jednotlivých záznamů v EThOS existují následující možnosti dostupnosti plných textů (38), podle volby instituce zpřístupňující práci:

1. okamžité stažení elektronické verze bez poplatku,
2. odkaz do repozitáře instituce přidělující titul,
3. digitální verze není dostupná, ale je možné objednat digitalizaci,
4. digitální verze není dostupná, uživatel musí kontaktovat instituci přidělující titul a dohodnout způsob zpřístupnění.

V případě, kdy práce není dostupná v elektronické podobě a první uživatel si objedná její digitalizaci, instituce vlastnící tištěnou verzi zasílá práci Britské knihovně, která ji za cenu nákladů digitalizuje. Po zaplacení (zájemcem nebo institucí, dle volby instituce) ceny digitalizace ve výši £43,29 + DPH, práce je v elektronické verzi zpřístupněna žadateli a zároveň bez poplatku všem dalším uživatelům EThOS. Uživatelé si mohou kromě stažení práce online objednat za poplatek práci vypálenou na CD/DVD, vytištěnou v kroužkové, měkké nebo tuhé vazbě. Na začátku projektu EThOS v roce 2009 bylo v rámci financování projektu Joint Information Systems Committee (zkráceně JISC) zdarma zdigitalizováno 10 000 prací, kvůli zajištění dostupnosti nejžádanějších prací online. Některé instituce v současné době v rámci vlastního financování zpětně digitalizují množství disertačních prací, které jsou následně uvolněny do EThOS ke stažení.

DART-Europe E-theses Portal (DEEP)

Projekt DART-Europe (Digital Access to Research Theses – Europe) vznikl v roce 2005 a trval 18 měsíců. Zakládajícími univerzitami byly University College London (UCL), Trinity

College Dublin, Oxford University a Dartington College of Arts ve spolupráci s ProQuest (40). Cíli projektu bylo:

- vytvoření Open Access portálu s evropskými eVŠKP,
- vytvoření hostované služby pro repozitáře za účelem podpory institucí bez lokálního repozitáře,
- příprava doporučení pro správu eVŠKP, institucionální repozitáře a metadata eVŠKP,
- ukázka služby pro ochranu digitálních objektů – eVŠKP,
- identifikace obchodních modelů pro dlouhodobou udržitelnost portálu, založenou na službách s přidanou hodnotou.

Vzhledem ke krátkému trvání programu se nepodařilo naplnit všechny cíle (službu pro digitální ochranu objektů, obchodní modely), některé byly rozpracovány až po skončení projektu nebo v rámci jiných projektů (metadata v rámci projektů EThOS a DRIVER).

V současné době je DART-Europe partnerstvím výzkumných knihoven a konsorcií sdružených za účelem podpory globálního přístupu k evropským výzkumným kvalifikačním pracím. DART-Europe je podporováno Access Division asociace LIBER (Ligue des Bibliothèques Européennes de Recherche) a je evropskou pracovní skupinou NDLTD.

Programu DART-Europe se účastní 28 partnerů z řad univerzit, národních knihoven, konsorcií a výzkumných institucí. Zájemci o vstup do programu DART-Europe podepisují dohodu (41)[5], ve které se zavazují dodržovat sedm principů, a to především podporovat vytváření, objevování a využití evropských eVŠKP, přispívat metadaty do portálu DART-Europe, přispívat k provozu a vývoji DART-Europe projektu, prosazovat status DART-Europe jakožto mezinárodního střediska excelence informací, odborných znalostí a zdrojů souvisejících s eVŠKP. DART-Europe je financován z příspěvků partnerských institucí.

[5] Česká verze dohody (viz Příloha II DART – Europe Dohoda o partnerství) byla zpracována autorem disertační práce pro potřeby VŠE v Praze při přistupování VŠE v Praze do DART-Europe.

Nejvýznamnějším výstupem projektu je DART-Europe E-theses Portal (zkráceně DEEP), který v lednu 2015 indexoval na 570 000 vědeckých kvalifikačních prací[6] (převážně doktorských) z 570 univerzit z 28 evropských zemí, dostupných v režimu Open Access. Počet prací v DART-Europe E-theses Portal v členění podle roku obhajoby, tj. včetně retrospektivních prací, znázorňuje Graf 1. Data jsou platná k 22. únoru 2015.

Správcem portálu DEEP je University College London, reprezentovaná správcem digitálních sbírek Martinem Moylem, který na pozvání autora této disertace v roce 2008 v rámci 4. ročníku semináře Systémy pro zpřístupňování VŠKP prezentoval možnosti zapojení České republiky do portálu DEEP (42).

Graf 1 Počet prací podle data obhajoby v portálu DEEP (zdroj: autor)

Vyhledávací rozhraní DEEP umožňuje jak jednoduché vyhledávání, tak prohlížení podle univerzity, zdroje dat (repozitáře nebo agregátora, zprostředkovatele dat), země nebo roku. Při fulltextovém vyhledávání je možné použít booleovské operátory AND (výchozí), OR a NOT, pravostranné rozšíření znakem *, uvozovky pro vyhledání frází a závorky pro zpřesnění dotazu. Výsledkovou množinu je možné zpřesnit pomocí faset země, roku, příjmení autora, zdroje dat, univerzity a jazyka práce. Funkci objevování zajímavých kvalifikačních prací

[6] V portálu DART-Europe je, kromě prací VŠE v Praze, zaindexována jedna práce s místem obhajoby Česká republika, práce je psána v německém jazyce.

(discovery) podporují na hlavní stránce seznamy posledních odevzdaných prací a nejvíce stahovaných prací za uplynulé období (viz Obrázek 4).

Obrázek 4 Vyhledávací rozhraní DART-Europe E-theses Portal (43)

Dříve používané pokročilé vyhledávání s možností upřesnit vyhledávání podle názvu, autora, roku obhajoby, univerzity, jazyka, klíčových slov a slov z abstraktu bylo z portálu odstraněno.

Detailní záznam eVŠKP obsahuje název práce, jméno autora, abstrakt, typ a formát práce, rok obhajoby a identifikátory – URL odkaz na plný text, citaci a URL odkaz na metadatový záznam. Vybrané záznamy je možné sdílet – odeslat na zvolený e-mail – a uložit pro pozdější použití do seznamu Marked List.

Import metadat probíhá protokolem OAI-PMH za využití metadatového standardu Dublin Core. Přispívat mohou libovolní agregátoři nebo instituce s otevřeným přístupem k eVŠKP. V rámci disertační práce byly portálem DEEP zaindexovány disertační práce Vysoké školy ekonomické v Praze, viz oddíl 4.3.3 *Realizace exportu na VŠE*.

Historií vzniku a kritickým hodnocením portálu DEEP se detailněji zabývá ve svém článku Martin Lochman na webu Ikaros, který za největší slabinu systému považuje „absenci jakéhokoliv jednotného systému věcného pořádání, která znemožňuje tematické oborové vyhledávání. Jde o neustále se prohlubující problém, který výraznou měrou snižuje efektivitu vyhledávání v systému.“ (44).

2.3.2 České repozitáře

Digitální repozitáře eVŠKP nabyly v České republice na důležitosti především po novele vysokoškolského zákona platné od 1. ledna 2006, která vysokým školám ukládala povinnost nevýdělečně zveřejňovat „disertační, diplomové, bakalářské a rigorózní práce, u kterých proběhla obhajoba, včetně posudků oponentů a výsledku obhajoby prostřednictvím databáze kvalifikačních prací“ (5 § 47b), kterou spravují, přičemž konkrétní způsob zveřejnění byl ponechán na samotné škole. Vznikla tak řada individuálních systémů na jednotlivých školách s odlišným způsobem provozu.

Níže si představíme charakteristiky tří vybraných systémů provozovaných na univerzitách v ČR a dvou agregátorů eVŠKP.

Systémy na bázi DSpace

V České republice byl v roce 2009 podle průzkumu (20) přibližně na třetině škol používán software DSpace. Jeho obliba v ČR je dána především velkou komunitou vývojářů DSpace na českých vysokých školách, soustředěných v diskusní skupině DSpaceCZ

(https://mailman.muni.cz/mailman/listinfo/dspacecz). V roce 2015 je repozitář eVŠKP na bázi DSpace používán na těchto školách:

- Akademie múzických umění v Praze
- České vysoké učení technické v Praze
- Technická univerzita v Liberci
- Univerzita Pardubice
- Univerzita Tomáše Bati ve Zlíně
- Vysoká škola báňská - Technická univerzita Ostrava
- Vysoké učení technické v Brně
- Západočeská univerzita v Plzni[7]

Systém DSpace ukládá metadata ve formátu Dublin Core, pro evidenci eVŠKP je potřeba rozšíření metadatového formátu o další prvky metadatového standardu EVSKP-MS (1). Někteří provozovatelé to řeší zavedením vlastních metadatových prvků, např. na Univerzitě Pardubice prvkem dc.thesis.degree-discipline pro studijní program a studijní obor eVŠKP. Záznamy jsou členěny do komunit a kolekcí, většinou podle fakult a typu prací.

Za referenční implementací EVSKP-MS v DSpace lze považovat digitální repozitář Akademie múzických umění v Praze - AMU (http://dspace.amu.cz/) obsahující metadata všech eVŠKP od roku 2006 (část prací od roku 2001) a plná znění většiny textových prací ve formátu PDF. Export metadat eVŠKP je řešen protokolem OAI-PMH s podporou EVSKP-MS. Vzhledem k tomu, že na AMU se obhajují také VŠKP s charakterem netextových prací (např. přednesení skladby), po konzultacích s autorem disertační práce bylo na AMU připraveno a realizováno referenční řešení správy netextových typů VŠKP v DSpace využívající vzájemné provázání souvisejících záznamů (např. textová práce a filmový záznam). Pokud student v licenci udělené škole nesouhlasil s užitím práce, plné texty nejsou veřejnosti v repozitáři AMU dostupné.

Na AMU je možné metadatové záznamy fulltextově prohledávat podle jednotlivých polí nebo procházet podle kolekce (fakulty), předmětu, názvu, jména autora nebo data vytvoření VŠKP,

[7] Kolekce VŠKP v DSpace ZČU obsahuje poslední příspěvek v roce 2013. Univerzita nyní pro VŠKP primárně používá informační systém školy IS/STAG.

seřadit podle názvu, data zaslání nebo data vytvoření VŠKP. Systém DSpace neumožňuje další zpřesňování dotazu např. pomocí faset.

Databáze kvalifikačních prací VŠE

Koncem března 2006 byl na podnět knihovny CIKS ustanoven prorektorem pro studijní a pedagogickou činnost VŠE v Praze odborný tým odpovědný za výběr a implementaci systému elektronického sběru dat a přípravu podkladů pro vnitřní předpisy školy. Tým byl složen z následujících členů:

- vedoucí týmu - člen Komise eVŠKP Jan Mach (návrh repozitáře, koordinace týmu)
- zástupce knihovny (integrace metadat VŠKP do knihovního systému Aleph)
- zástupce Výpočetního centra (instalace aplikace, integrace s informačním systémem VŠE v Praze)
- děkan fakulty informatiky a statistiky (komunikace s vedením fakult, implementace workflow)

Úkolem týmu bylo specifikovat workflow pro potřeby VŠE v Praze, připravit vzory předpisů, příprava dat a řešení koncepce převodu dokumentů do PDF jakožto hlavního formátu pro odevzdávání VŠKP. Pro potřeby VŠE v Praze byla připravena Databáze kvalifikačních prací VŠE (45), spuštěná v květnu 2006 na adrese http://www.vse.cz/vskp. Aplikace je naprogramována v jazyce PHP, provozována na webovém serveru Apache s databází MySQL. VŠE v Praze se tak stala jednou z prvních škol, na které byl vybudován repozitář eVŠKP podle *Souboru doporučení* (46) Komise eVŠKP.

Na základě podkladů pracovního týmu byl vydán Pokyn prorektora pro studijní a pedagogickou činnost upravující odevzdávání a zpřístupňování eVŠKP na VŠE v Praze. Repozitář obsahuje metadata podle standardu EVSKP-MS (1), plné texty prací, posudky a přílohy. Utajované informace mohou být po odsouhlasení uloženy zvlášť v příloze, která není veřejnosti přístupná.

Díky implementovanému protokolu OAI-PMH bylo v roce 2008 možné server použít pro export dat v rámci projektu Theses.cz, později např. i pro export metadat do Národního úložiště šedé literatury a dalších služeb.

Vzhledem ke stanovisku Samostatného oddělení autorského práva Ministerstva kultury (47) plné texty prací zprvu nebyly volně přístupné na Internetu. Pro získání přístupu se zájemci mimo akademickou obec VŠE v Praze museli osobně zaregistrovat v knihovně CIKS. Od září 2011, po novelách Autorského zákona a na základě konzultací s právníky v rámci zpracovávaného projektu NUŠL (viz oddíl 2.1.3 Zveřejňování a sdělování), začala VŠE v Praze zpřístupňovat plné texty volně na Internetu bez nutnosti registrace. Student uděluje souhlas se sdělováním díla volně na Internetu odevzdáním práce do informačního systému (konkludentní uzavření licenční smlouvy). Zveřejněním plných textů eVŠKP je tak naplněna povinnost daná škole VŠ zákonem 111/1998 Sb. v § 47b odst. 1.

V rámci repozitáře autor disertační práce navrhl OAI-PMH server pro export metadat eVŠKP ve standardu EVSKP-MS a metadat článků z odborných časopisů VŠE v Praze. Metadata harvestují např. služby Theses.cz, NUŠL, PRIMO, DART-Europe E-theses Portal aj. Případová studie implementace OAI-PMH je samostatně popsána v podkapitole 4.3.

Po zavedení nového studijního informačního systému ISIS byla agenda odevzdávání eVŠKP převedena do informačního systému, ze kterého jsou metadata přebírána do Databáze kvalifikačních prací VŠE a knihovního katalogu Aleph. Databáze kvalifikačních prací VŠE po mírných úpravách dodnes slouží pro účely vyhledávání eVŠKP VŠE v Praze a jako OAI-PMH server pro export metadat.

Vyhledávací rozhraní nad databází je poplatné době vzniku, vyhledávací formulář umožňuje pouze fulltextové vyhledávání v základních metadatech (název, autor, abstrakt) a filtrování podle studijního programu/oboru, roku a typu práce (viz Obrázek 5). V kapitole 7 disertační práce je proto řešen potřebný redesign vyhledávacího rozhraní.

Obrázek 5 Filtr seznamu v Databázi kvalifikačních prací VŠE (45)

Repozitář závěrečných prací Univerzity Karlovy v Praze

Repozitář závěrečných prací Univerzity Karlovy v Praze (zkráceně UK) je dostupný na adrese https://is.cuni.cz/webapps/zzp . Navazuje na zpřístupňování VŠKP prostřednictvím tzv. věcné databáze kvalifikačních prací (VŠKP umístěné zpravidla v knihovně příslušné fakulty) a prostřednictvím repozitáře v systému DigiTool (24).

Repozitář závěrečných prací UK obsahuje obhájené bakalářské, diplomové, rigorózní a disertační práce od roku 2006 (část dat migrována z Aleph a Digitool). Práce v repozitáři UK jsou kontrolovány na podobnost textu v rámci projektu Theses.cz. Práce jsou volně dostupné veřejnosti na základě § 47b zákona o vysokých školách, v souladu s principy Open Access[8]. Citlivé údaje mají studenti možnost uvést v příloze, která se se souhlasem děkana (příp. proděkana) nezveřejní. Výjimečně lze se souhlasem rektora nezveřejnit celou práci.

[8] Práce původně dostupné v systému DigiTool byly přístupné pouze z IP adres univerzity, akademické obci Univerzity Karlovy, nebo po registraci.

Zveřejňování prací v repozitáři UK je v souladu s čl. 18a *Studijního a zkušebního řádu UK* a čl. 6a *Rigorózního řádu UK* a *Opatřením rektora č. 6/2010* ve znění *Opatření rektora č. 14/2014*. Postup evidence a zveřejňování eVŠKP na UK, včetně popisu workflow a odpovědných osob, je upraven metodickým pokynem (48) vycházejícím z *Opatření rektora č. 6/2010*. Oproti dřívějším systémům zpřístupňování eVŠKP na UK (24), práce v současném repozitáři jsou dostupné veřejně bez omezení IP adres nebo podmínky registrace.

Uživatelské rozhraní umožňuje fulltextové prohledávání metadat na zadaná klíčová slova, ve výchozím nastavení spojená booleovským operátorem OR. Vyhledávání nerozlišuje velká a malá písmena, diakritiku. Vyhledávací systém podporuje vyhledávání frází (použití uvozovek), operátorů AND a OR, operátoru vyloučení (znak mínus, NOT) anebo zahrnutí (znak plus), vyhledání frází s uživatelsky definovanou vzdáleností slov (např. "Karel Mácha"~1) a upravování váhy slova v dotazu (^číslo).

Uživatelské rozhraní obsahuje fasetový navigační systém, který umožňuje zúžit seznam nalezených eVŠKP podle typu práce, fakulty, roku obhajoby, typu dokumentu anebo jazyka práce. Fasety neobsahují informaci o četnosti záznamů v rámci subkategorií dané fasety, uživatel tak předem nemůže odhadnout, jak moc bude seznam zúžen.

Národní registr VŠKP – Theses.cz

Metadatové záznamy a plné texty eVŠKP byly před realizací tzv. národního registru VŠKP, Theses.cz, přístupné veřejnosti pouze v lokálních databázích vysokých škol. Původní návrh centrálního národního registru VŠKP vznikl v rámci Komise eVŠKP na základě rozsáhlé diskuse mezi zapojenými školami. Vyústil přípravou centralizovaného rozvojového projektu *Národní registr VŠKP* pro rok 2007 pod vedením Vysoké školy ekonomické v Praze[9]. Centralizovaný projekt nebyl z důvodu nedostatku financí MŠMT ČR přijat, proto byl v obdobném znění řešiteli připraven i na rok následující.

Projekt *Odhalování plagiátů v závěrečných pracích* na rok 2008, odpovídající projektu *Národní registr VŠKP*, podala i Masarykova univerzita, která se na základě předchozích žádostí univerzit rozhodla přizpůsobit svůj interní antiplagiátorský systém potřebám vysokých

[9] Zřizovatel Komise eVŠKP – Asociace knihoven vysokých škol ČR – nemůže samostatně podávat rozvojové projekty, neboť není vysokou školou dle zákona.

škol. Součástí projektu MUNI měl být i digitální repozitář eVŠKP, který by sloužil jako zdroj dokumentů pro potřeby vyhledávání duplicit.

Po výzvě České konference rektorů ze dne 6. 9. 2007 se uskutečnilo společné jednání řešitelů projektů *Národní registr VŠKP* (koordinátor VŠE v Praze, Jan Mach) a *Odhalování plagiátů v závěrečných pracích* (koordinátor MUNI, Michal Brandejs). Cílem jednání bylo najít společné řešení obou projektů z důvodů jejich obsahové podobnosti, resp. možnost jejich sloučení.

Na základě výše uvedeného setkání bylo vydáno prohlášení řešitelů (49) a následně připraven společný centralizovaný rozvojový projekt *Národní registr vysokoškolských kvalifikačních prací – metadata, plné texty a řešení projevů plagiátorství*. Projekt byl podán v rámci programu 3. *Program na rozvoj přístrojového vybavení a moderních technologií*, podprogram b) *rozvoj informačních a komunikačních technologií (včetně multilicencí softwarových produktů)*. Projektu na vybudování národního registru VŠKP se účastnilo následujících 17 škol (50):

- Akademie múzických umění v Praze
- Česká zemědělská univerzita v Praze
- Janáčkova akademie múzických umění v Brně
- Jihočeská univerzita v Českých Budějovicích
- Masarykova univerzita
- Ostravská univerzita v Ostravě
- Slezská univerzita v Opavě
- Univerzita Hradec Králové
- Univerzita Jana Evangelisty Purkyně v Ústí nad Labem
- Univerzita Palackého v Olomouci
- Univerzita Tomáše Bati ve Zlíně
- Vysoká škola báňská – Technická univerzita Ostrava
- Vysoká škola ekonomická v Praze
- Vysoká škola polytechnická Jihlava
- Vysoká škola technická a ekonomická v Českých Budějovicích
- Vysoká škola uměleckoprůmyslová v Praze
- Západočeská univerzita v Plzni

Hlavním řešitelem rozvojového projektu byl Michal Brandejs (MUNI), koordinací návrhu struktur a funkcí národního registru VŠKP byl ustanoven Jan Mach (VŠE v Praze). Projekt navázal na práci Komise eVŠKP (využití analýz, metodik, doporučení a vzorů pro eVŠKP) a na širší projekt tehdejší Státní technické knihovny v oblasti Národního uložiště šedé

literatury (spoluřešitel Jan Mach, VŠE v Praze), jehož nedílnou složkou se národní registr VŠKP měl stát.

Primárními cíli projektu bylo:

- vytvoření národního registru (metadata),
- hledání podobných textů v úložišti eVŠKP,
- možnost zpřístupňování eVŠKP dle rozhodnutí školy.

V dílčím projektu za VŠE v Praze byly pro rok 2008 stanoveny následující cíle:

- specifikace systému Theses.cz,
- metadatový standard VŠKP pro import/export,
- koordinace přípravy Theses.cz, připomínkování, testy,
- export záznamů z lokálního registru VŠE v Praze protokolem OAI-PMH (viz oddíl 4.3.3),
- prezentace projektu, nástrojů a dalších materiálů odborné veřejnosti na 3. ročníku semináře Systémy pro zpřístupňování eVŠKP,
- spuštění národního registru vysokoškolských kvalifikačních prací Theses.cz.

V rámci projektu byly implementovány následující funkce Theses.cz:

- shromažďování, zveřejňování metadat a jejich správa,
- fulltextové a tematické vyhledávání záznamů,
- zpřístupnění plných textů dle nastavení školy,
- sběr plných textů pro potřeby vyhledávání podobností,
- vyhledávání podobností.

Provoz systému Theses.cz je řešen na základě dvoustranných smluv, mezi poskytovatelem služby (zpracovatelem dat – Masarykovou univerzitou) a mezi poskytovatelem dat (spolupracující školou). Odlišné požadavky a práva na zpracování a zpřístupnění metadat a plných textů jsou řešeny v tzv. konfigurátoru Theses.cz, nastavení systému specifické pro školu je součástí smluv mezi MUNI a spolupracující školou.

Řešitelský tým MUNI v rámci implementace projektu odmítl zpřístupnit agregovaná metadata s tím, že k tomu nemá souhlas jednotlivých univerzit. Ze stejných důvodů nebylo možné dojednat exportování metadat do Národního úložiště šedé literatury a nadnárodního registru DEEP (viz oddíl 2.3.1, část *DART-Europe E-theses Portal (DEEP)*). I přes zájem univerzit o možnost exportu dat z Theses.cz však řešitelé do současné doby nepožádali zapojené

univerzity o oficiální souhlas a neumožňují export metadat a příp. plných textů z Theses.cz, jak bylo původně přislíbeno v prohlášení řešitelů[10].

Na základě dotazníku Komise eVŠKP (20) víme, že do října 2009 předalo metadata do Theses.cz osm z deseti respondentů. 47 % respondentů využívalo možnosti dávkového zasílání, 38 % respondentů nahrávalo metadata ručně a pouze 15 % zpřístupňovalo metadata standardem OAI-PMH. 60 % škol používalo pro export formát Theses.cz, zbylých 40 % metadatový standard EVSKP-MS. Pouze u tří škol byly metadatové záznamy doplněny odkazem na plný text práce v lokálním úložišti školy. Aktualizace stavu zpřístupňování eVŠKP v Theses.cz v roce 2014 je součástí zadání průzkumu v kapitole 3.

Na projekt Theses.cz navázala řada dalších centralizovaných rozvojových projektů MŠMT ČR, které rozvíjely projekt první. Vznikly tak systémy související s Theses.cz – systém Odevzdej.cz zaměřený na detekci plagiátorství v seminárních prací a systém Repozitar.cz pro zpřístupnění a detekci plagiátorství v odborných vědeckých pracích. VŠE v Praze, spoluřešitel projektů, v současné době používá vlastní rozhraní do uvedených systémů http://validator.vse.cz. Implementace projektu Validátor VŠE je blíže popsána v rámci kapitoly 9.

Autor disertační práce během působení v Komisi eVŠKP opakovaně poukazoval na omezení, které má základní i pokročilé uživatelské rozhraní Theses.cz.

V případě základního uživatelského rozhraní Theses.cz byl, a doposud v jisté formě i nadále je, problém v použité syntaxi vyhledávání, která neodpovídá zvyklostem uživatelů Internetu. Původně se zadaný text vyhledával jako fráze, až později bylo použito běžnější syntaxe spojující ve výchozím stavu jednotlivá slova operátorem AND.

[10] „V rámci projektu bude připraveno otevřené rozhraní pro napojení dalších systémů třetích stran, které by umožnily vyhledávání plagiátů v plných textech v rámci národního registru.“ (47)

Podrobná nápověda není primárně určena pro Theses.cz, ale pro informační systém Masarykovy univerzity[11]. Funkčnost booleovských operátorů v rozhraní proto autor opakovaně ověřoval empirickým pokusem, naposledy v březnu 2015.

Základní vyhledávací rozhraní (https://Theses.cz/vyhledavani/) podle testů spojuje jednotlivé termíny ve vyhledávání booleovským operátorem AND. Vyloučení termínu je umožněno použitím znaku - (mínus) před termínem. Slova AND, OR a NOT jsou interpretována v dotazu jako vyhledávané termíny, nikoliv jako booleovské operátory. Při nalezení více jak 1 000 záznamů, textové informační hlášení o počtu nalezených záznamů chybně udává jako počet nalezených záznamů vždy číslo menší než 1 000. Chybná informace o počtu nalezených záznamů může souviset s urychlením odezvy vyhledávání v poslední verzi Theses.cz nasazené na podzim 2014 (původní doba odezvy vyhledávání byla v řádu jednotek sekund).

V případě tzv. *katalogového hledání* Theses.cz (https://Theses.cz/th_search/tematicke.pl) je k dispozici filtr s výběrem polí Pracoviště závěrečné práce, Příjmení (resp. první dvě písmena), Rok a Titul. Na rozdíl od základního rozhraní není možné omezit vyhledávání podle zadaného slova nebo fráze. Pokud vybereme kombinaci hodnot v jednotlivých polích takovou, kdy zadané podmínce nevyhovuje ani jeden záznam, nevíme, které pole by bylo vhodné upravit pro získání nenulového výsledku. Již nalezené záznamy není možné dále zpřesňovat nebo třídit.

Národní úložiště šedé literatury

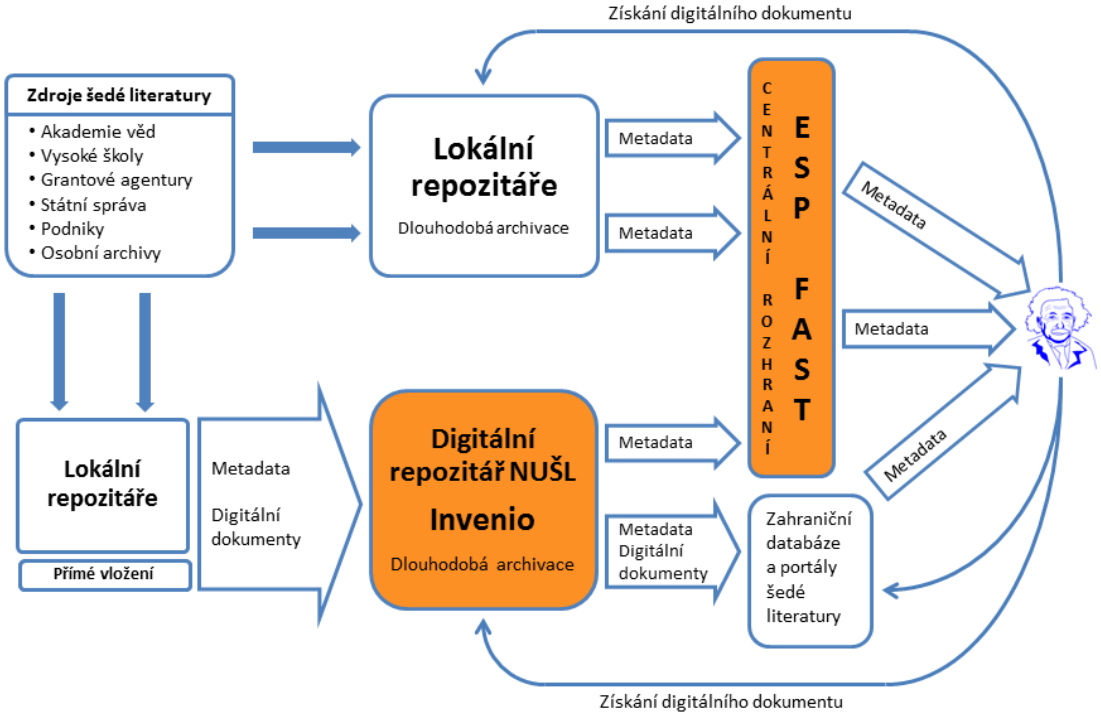
Národní úložiště šedé literatury (zkráceně NUŠL) poskytuje centrální přístup k informacím o šedé literatuře, vznikající v České republice v oblastech vědy, výzkumu a vzdělávání. Národní úložiště šedé literatury bylo vybudováno díky podpoře Ministerstva kultury ČR v rámci projektu *Digitální knihovna pro šedou literaturu - funkční model a pilotní realizace*. Jednalo se o čtyřletý projekt (2008 – 2011), který řešila Národní technická knihovna (zkráceně NTK) společně s Vysokou školou ekonomickou v Praze (řešitel Jan Mach).

[11] Systém Theses.cz vychází z kódu informačního systému Masarykovy university. Podrobná nápověda pro IS MUNI uvedená v Theses.cz (https://theses.cz/help/komunikace/vyhledavani) tak obsahuje i irelevantní informace, např. omezení vyhledávání na určitou agendu.

Součástí řešení projektu byla příprava metadatového formátu přímo pro potřeby NUŠL. Hlavní požadavky na formát byly jednoduchost, minimum povinných polí a respektování metadatového standardu Dublin Core. Vysokoškolské kvalifikační práce jsou jedním ze základních druhů šedé literatury, proto metadatový formát NUŠL používá prvky z formátů Dublin Core, EVSKP-MS, ETD-MS a vlastní prvky. Formát NUŠL byl zpracován v roce 2008, následně byl podroben odborné expertize a testován na vlastních datech NTK a VŠE v Praze. Vzhledem k tomu, že systém Invenio nativně využívá metadatový formát MARC 21, byla zpracována konverzní tabulka mapující prvky formátu NUŠL na MARC 21 (51). Obdobné mapování je provedeno autorem disertační práce pro formát EVSKP-MS v podkapitole 4.2 *Mapování prvků formátu EVSKP-MS*.

V rámci projektu byla řešiteli provedena analýza open source software pro digitální knihovny (konkrétně DSpace, Fedora, Invenio, Eprints a Greenston). Pro Digitální repozitář NUŠL (http://invenio.ntkcz.cz) byl vybrán software Invenio, pro Centrální rozhraní NUŠL (http://www.nusl.cz) software FAST ESP poskytující indexování a vyhledávání nad metadaty v Digitálním repozitáři NUŠL a v lokálních repozitářích zapojených institucí. Obrázek 6 znázorňuje schéma realizovaného softwarového řešení. Univerzity pro indexování eVŠKP nejčastěji využívají zpřístupnění metadat eVŠKP protokolem OAI-PMH. Prvky EVSKP-MS jsou mapovány na odpovídající prvky formátu NUŠL v rámci importu. Schéma dále znázorňuje zpřístupnění metadat z Digitálního repozitáře NUŠL do zahraničních repozitářů a portálů šedé literatury. Od roku 2014 jsou metadata NUŠL, včetně zaindexovaných eVŠKP zapojených institucí, zpřístupněny v rámci repozitáře OpenAIRE (https://www.openaire.eu/).

Centrální vyhledávací rozhraní na adrese http://www.nusl.cz poskytuje možnost využití faset Zdroj (v horním menu), Typ dokumentu (včetně omezení na VŠKP), Osoby, Klíčová slova, Jazyk a dostupnost plného textu. Při fulltextovém vyhledávání jsou vyhledávané termíny spojeny booleovským operátorem AND, pokud není uvedeno v dotazu jinak.



Obrázek 6 Schéma softwarového řešení NUŠL (118)

2.4 Závěr kapitoly

V úvodu kapitoly autor definuje základní pojmy, které se týkají tématu disertační práce. Kromě definice VŠKP je stanoveno, jak interpretovat pojmy zpřístupňování a sdělování eVŠKP veřejnosti v souladu se zákonem a zveřejňování dokumentů v režimu Open Access.

V závěrečných doporučeních se autor přiklání k prezentovanému názoru, že školy mají povinnost danou zákonem o vysokých školách práce zpřístupňovat na Internetu tak, aby kdokoliv k nim měl přístup na místě a v čase podle své vlastní volby. Nelze přístup podmiňovat udělením souhlasu univerzitou, zpřístupněním pouze v prostorách školy apod. bariérami. Odevzdáním práce její autor uděluje souhlas se zpřístupněním eVŠKP na Internetu, licence je uzavřena konkludentně nahráním eVŠKP do repozitáře univerzity. Nedochází tak k rozporu s Autorským zákonem.

Poslední definice této kapitoly se zabývá plagiátorstvím. Problematika plagiátorství nesouvisí jen s odhalováním plagiátorství v odevzdávaných textech studentů, ale i s prevencí. Na Vysoké škole ekonomické v Praze proto zaměstnanci knihovny pravidelně pořádají přednášky zaměřené na citační etiku a správné použití citační normy *ČSN ISO 690 (01 0197)*. Studenti

také aktivně využívají WWW stránky knihovny http://ciks.vse.cz/citace s návody, s často kladenými otázkami k citacím a s možností zaslat specializovaným oborovým knihovníkům dotaz ke konkrétní problematice správného citování. Přednášky knihovny a podpora správné citační praxe jsou důležitou složkou prevence především neúmyslného plagiátorství. Kromě volby vhodné softwarové podpory vyhledávání plagiátů v eVŠKP, která bude předmětem kapitoly 8, autor knihovnám doporučuje zaměřit se na podporu správné citační praxe formou výukových materiálů, seminářů a konzultací.

Jak vyplynulo z analýzy vývoje správy a zpřístupňování eVŠKP v ČR v uplynulých letech, po ukončení činnosti Komise eVŠKP se již touto problematikou nikdo systematicky nezabývá. Předmětné oblasti disertační práce se věnují některé závěrečné práce, které většinou čerpají buď z dotazníkových šetření Komise eVŠKP, nebo realizují průzkumy zaměřené jen na omezený počet institucí (většinou na školu autora) či jsou omezeny tematickým záběrem.

Protože jedním z cílů disertační práce je poskytnout aktualizaci znalostí o stavu správy a zpřístupňování VŠKP v ČR, je nutné v rámci disertační práce provést nové dotazníkové šetření, které by poskytlo přehled o reálném stavu uvedené problematiky a na základě výsledků stanovit doporučení pro další praxi a výzkum. Provedený průzkum zpřístupňování eVŠKP v ČR v roce 2014 je uveden v kapitole následující.

Studie zahraničních a českých repozitářů eVŠKP poukazuje na rozdíl v dostupné funkcionalitě ve vyhledávacích službách eVŠKP v ČR a v zahraničí. Webová aplikace Theses.cz je využívána jako tzv. národní registr VŠKP. Autor popisuje vznik a vývoj této služby, na které se projektově podílel jako spoluřešitel za VŠE v Praze. Při návrhu repozitáře však nebyly programátorským týmem Theses.cz přijaty návrhy Komise eVŠKP na úpravu funkcionality vyhledávacího rozhraní, které dle zjištění autora disertační práce i v současné době vykazuje zásadní chyby v implementaci booleovské logiky při vyhledávání a odezva systému na vyhledávací dotaz je velmi pomalá. Katalogové vyhledávání nepodporuje

kombinaci fulltextového vyhledávání a omezení na vybrané kategorie. Uživatelské rozhraní také nepodporuje fasety pro zpřesňování dotazů.

Zahraniční repozitáře umožňují následující funkce, jejichž implementaci je vhodné zvážit při návrhu repozitáře eVŠKP a vyhledávacího rozhraní nad metadaty a plnými texty eVŠKP:

- fulltextové vyhledávání v plných textech,
- fasetová navigace namísto pokročilého vyhledávání,
- procházení databáze podle vybraných kategorií,
- import/export metadat protokolem OAI-PMH,
- možnost odeslání záznamu e-mailem nebo export do citačního manažeru,
- přehled nových přírůstků v repozitáři,
- přehled nejžádanějších záznamů za uplynulé období,
- statistiky přístupů.

Mnohé výše uvedené funkce bohužel české repozitáře eVŠKP často postrádají, obsahují většinou pouze základní vyhledávání, procházení podle základních kategorií či podle data přírůstku, případně pokročilé vyhledávání s možností nastavení filtrů dotazu podle vybraných polí záznamu. Příkladem, kdy použitá technologie a funkcionalita již neodpovídá velkému množství indexovaných záznamů, je např. Databáze kvalifikačních prací VŠE. Český repozitář s vhodně řešeným vyhledáváním je Národní úložiště šedé literatury, které podporuje využití faset pro zpřesnění dotazu, ale je zaměřen šířeji, na šedou literaturu. Vyšší komfort vyhledávání nabízejí také discovery služby, které se již začínají objevovat na českých univerzitách (detailněji viz kapitola 6 *Výběr systému centralizovaného vyhledávání*), i když vzhledem k obecnějšímu zaměření nemohou konkurovat pokročilým vyhledávacím rozhraním specializovaným na eVŠKP.

V kapitole 7 proto autor na příkladu Databáze kvalifikačních prací VŠE představuje připravenou modelovou aplikaci vývoje aplikace pro indexování plných textů eVŠKP, metadat ve formátu EVSKP-MS a souvisejícího vyhledávacího rozhraní.

3 Průzkum zpřístupňování vysokoškolských kvalifikačních prací v roce 2014

Na podzim 2014 autor disertační práce provedl průzkum, jehož cílem bylo zjistit aktuální stav zpřístupňování elektronických vysokoškolských kvalifikačních prací na veřejných vysokých školách (zkráceně VVŠ). Navázal především na obdobná dotazníková šetření, která v letech 2006 (52), 2007 (53) a 2009 (20) realizovala Odborná komise pro otázky elektronického zpřístupňování vysokoškolských kvalifikačních prací AKVŠ ČR (Komise eVŠKP). Dále byly zohledněny dílčí výzkumy zpřístupňování eVŠKP v řešených bakalářských a diplomových pracích, které se však většinou týkaly pouze stavu na vybraných školách nebo se odvolávaly na výše zmíněné dotazníky Komise eVŠKP (viz podkapitola 2.2).

Text kapitoly vychází z článku publikovaného autorem disertační práce v časopise *ProInflow 2/2014* (19).

3.1 Výzkumná otázka

Základní výzkumná otázka, na kterou autor ve výzkumu hledal odpověď, byla „Jakým způsobem se zpřístupňují metadata a plné texty eVŠKP?“. Podotázky výzkumu byly:

1) Jak jsou metadata a plné texty eVŠKP sbírány a evidovány v repozitářích škol?
2) Jakým způsobem a kde jsou metadata a plné texty eVŠKP zpřístupňovány?
3) Jak je řešena problematika plagiátorství u eVŠKP?
4) Na základě jaké legislativy je řešeno zpřístupňování eVŠKP?
5) K jakým změnám došlo od posledního průzkumu Komise eVŠKP, lze vysledovat nový trend např. u použitých metadat, zpřístupňování eVŠKP a v utajování informací?

3.2 Příprava průzkumu

Ve snaze o plné pochopení stavu zpřístupňování eVŠKP v ČR bylo nutné přistoupit ke kombinaci kvalitativního a kvantitativního výzkumu. Základní šetření probíhalo formou strukturovaných dotazníků, které byly adresně zaslány vytipovaným správcům repozitářů eVŠKP.

Na základě výše uvedených otázek a získaných zkušeností z dotazníků eVŠKP byla formulována sada otázek a uzavřených, příp. polouzavřených odpovědí doplněná prostorem pro volný komentář. Použití uzavřených odpovědí umožnilo jednodušší vyhodnocení, polouzavřené otázky poskytly prostor popisu v nestandardní situaci. U některých otázek byla možná volba jedné odpovědi z mnoha, u jiných bylo možné označit více odpovědí správně, případně byl ponechán prostor pro volnou odpověď.

Abychom mohli sledovat vývojové trendy, bylo žádoucí ponechat otázky ve znění již realizovaných průzkumů nebo příp. je lehce upravit na základě poskytnuté zpětné vazby z již realizovaných průzkumů. Vzhledem k pětiletému časovému odstupu a posunu zkoumaných témat však bylo možné ponechat beze změn jen pár otázek.

V úvodu dotazníku byli respondenti seznámeni s kontaktem na realizátora výzkumu, pokyny a termínem pro vyplnění, cíli a záměry využití průzkumu a s použitými zkratkami.

Získané odpovědi byly následně doplněny a v případě potřeby verifikovány formou ad hoc interview s respondenty, evaluací repozitářů jednotlivých škol, národního registru VŠKP Theses.cz (http://www.theses.cz) a Národního úložiště šedé literatury (http://www.nusl.cz).

Příloha III obsahuje dotazník zasílaný respondentům.

3.3 Cílová skupina a respondenti

U popisného výzkumu je kladen velký význam na opakovaný sběr v delším časovém intervalu. Abychom mohli objektivně navázat na již realizované průzkumy Komise eVŠKP, respondenty dotazníkového šetření byly pouze veřejné vysoké školy v ČR.

Seznam všech 26 veřejných vysokých škol byl převzat ze stránek MŠMT ČR (54). Jako primární zdroj kontaktů k oslovení byl zvolen seznam správců repozitářů v systému Theses.cz (55), následně doplněný o osobní kontakty autora průzkumu z doby působení jako předseda Komise eVŠKP a jako řešitel projektů NUŠL, Theses.cz a následujících. U dalších škol byla vhodná kontaktní osoba vytipována mezi účastníky konference Otevřené repozitáře 2014 nebo z webu příslušné školy. Respondenty byli nejčastěji správci informačního systému školy, případně ředitelé či pracovníci výpočetních a informačních center nebo knihovny univerzity (kompetencemi útvarů za provoz repozitáře eVŠKP se zabývala samostatná otázka, viz Graf 3).

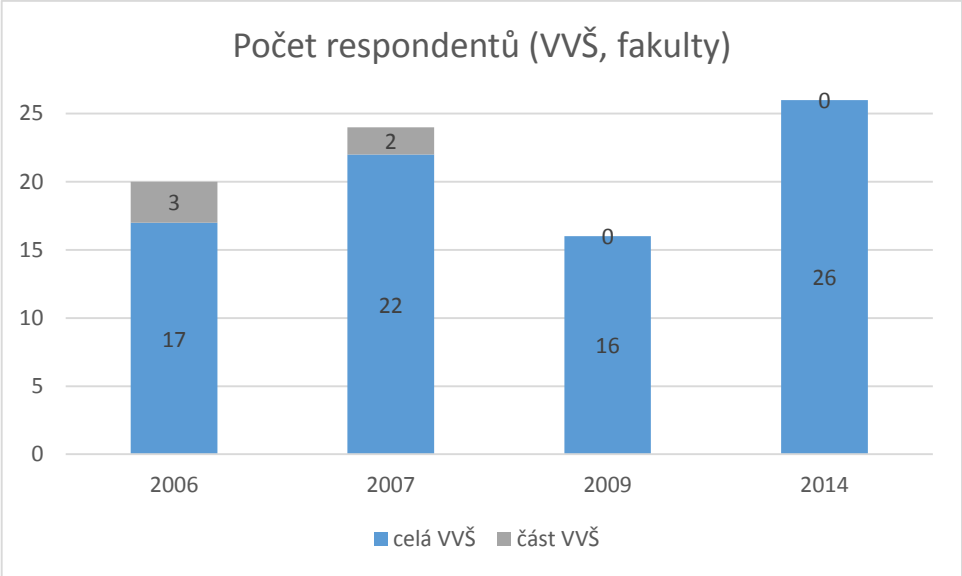
Rozeslání dotazníků zvoleným kontaktním osobám probíhalo v několika vlnách na přelomu srpna a září 2014, aby mohly být odladěny případné chyby nebo nejasnosti z prvních odpovědí. V případě, že ve stanovené lhůtě nebyla respondentem zaslána odpověď nebo pokud kontaktní osoba odmítla vyplnění dotazníku[12], autor průzkumu se snažil najít další kontakt na škole, což se dařilo většinou díky osobnímu rozhovoru s pracovníky školy a především zásluhou zaměstnanců knihoven na dané vysoké škole.

Podařilo se získat vyplněný dotazník od všech 26 oslovených veřejných vysokých škol[13], pouze výjimečně bylo potřeba až měsíční jednání za účelem získání odpovědí. Počet respondentů ve srovnání s předchozími průzkumy Komise eVŠKP je uveden v Grafu 2.

V průzkumu v roce 2006 odpovědělo 17 veřejných vysokých škol jako celek, vybrané fakulty 3 univerzit, 1 státní škola; sumarizace byla nejčastěji uváděna za 17 škol, příp. doplněná o trendy z fakult Univerzity Karlovy a ze státní školy. V roce 2007 narostl počet respondentů na úrovni celé veřejné vysoké školy na 22 plus dílčí odpovědi po fakultách z Univerzity Karlovy a ze Slezské univerzity. V roce 2009 byly získány odpovědi pouze z 16 veřejných vysokých škol, přičemž Univerzita Karlova odpovídala vzhledem k roztříštěnosti řešení problematiky zpřístupňování eVŠKP opět po fakultách.

[12] Odmítnutí odpovědí bylo nejčastěji z důvodu pracovní zaneprázdněnosti na začátku září, např. při zavádění nového informačního systému. Veterinární a farmaceutická univerzita odmítla poskytnut odpovědi en block kvůli tématu dotazníku, zde se autorovi podařilo získat odpovědi z veřejných zdrojů a na základě telefonního rozhovoru s jednotlivými pracovníky školy.

[13] Seznam škol: Akademie múzických umění v Praze (AMU), Akademie výtvarných umění v Praze (AVU), Česká zemědělská univerzita v Praze (ČZU), České vysoké učení technické v Praze (ČVUT), Janáčkova akademie múzických umění (JAMU), Jihočeská univerzita v ČB (JU), Masarykova univerzita (MU), Mendelova univerzita v Brně (MENDELU), Ostravská univerzita v Ostravě (OU), Slezská univerzita v Opavě (SU), Technická univerzita v Liberci (TUL), Univerzita Hradec Králové (UHK), Univerzita Jana Evangelisty Purkyně (UJEP), Univerzita Karlova v Praze (UK), Univerzita Palackého v Olomouci (UPOL), Univerzita Pardubice (UPA), Univerzita Tomáše Bati ve Zlíně (UTB), Veterinární a farmaceutická univerzita (VFU), Vysoká škola báňská - Technická univerzita Ostrava (VSB-TUO), Vysoká škola ekonomická v Praze (VŠE), Vysoká škola chemicko-technologická (VŠCHT), Vysoká škola polytechnická Jihlava (VŠPJ), Vysoká škola technická a ekonomická (VŠTE), Vysoká škola uměleckoprůmyslová (VŠUP), Vysoké učení technické v Brně (VUT), Západočeská univerzita v Plzni (ZČU)



Graf 2 Počet respondentů (veřejné vysoké školy, fakulty) (zdroj: autor)

Oproti předchozím rokům se v roce 2014 poprvé podařilo získat odpovědi od všech veřejných vysokých škol. Odpovědi byly získány uceleně za celou vysokou školu, a to i v případě Univerzity Karlovy, která již problematiku zpřístupňování eVŠKP celkem zásadněji sjednotila na všech fakultách (viz oddíl 2.3.2). Poprvé v historii těchto průzkumů byly získány odpovědi od škol:

- Vysoká škola polytechnická Jihlava,
- Vysoká škola technická a ekonomická,
- Vysoká škola uměleckoprůmyslová.

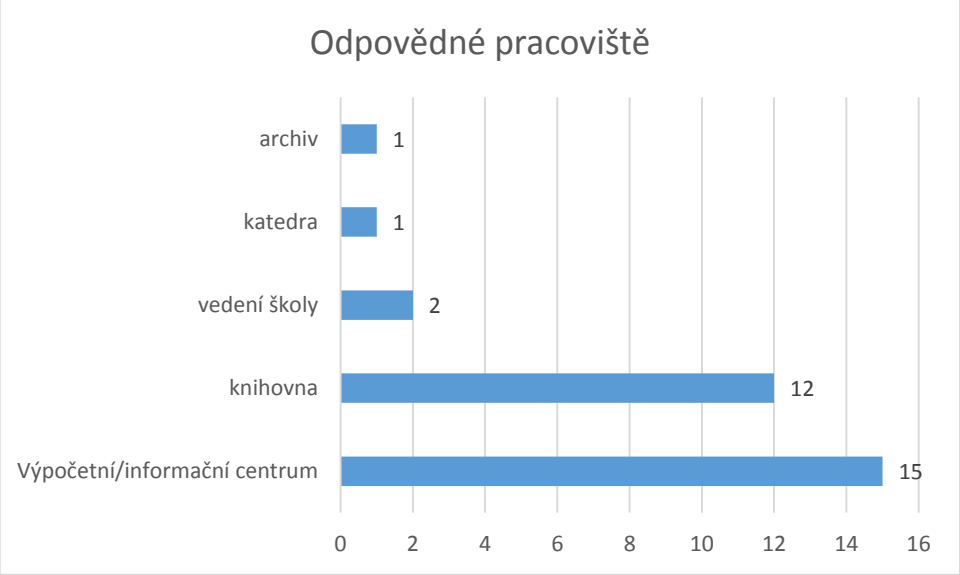
V průzkumech nebyla řešena problematika zpřístupňování eVŠKP na soukromých vysokých školách, jejichž počet stále v ČR narůstá. Stav na soukromých vysokých školách by si proto zasloužil samostatný výzkum, který je nad rámec této disertační práce.

3.4 Vyhodnocení jednotlivých dotazů

Níže jsou uvedeny souhrnné výsledky jednotlivých dotazů, vyhodnocení a případné srovnání s předchozími průzkumy Komise eVŠKP. Vzhledem k přislíbení anonymity u odpovědí jsou uvedeny konkrétní názvy škol pouze tam, kde autor získal od respondenta souhlas se zveřejněním jména nebo se jedná o veřejně dostupnou informaci. V grafech jsou uvedeny údaje tak, jak byly vyplněny respondenty. Pokud autor zjistil rozpor se skutečností, je na to upozorněno v textu.

3.4.1 Úroveň koordinace problematiky zpřístupňování eVŠKP

Útvar na škole pověřený řešením problematiky zpřístupňování eVŠKP bylo možné v dotazníku vypsat volným textem (viz Graf 3). V pěti případech byla uvedena kombinace více útvarů, z toho 4× kombinace výpočetního/informačního centra a knihovny. Pouze na jedné škole je problematika řešena decentralizovaně na fakultách.

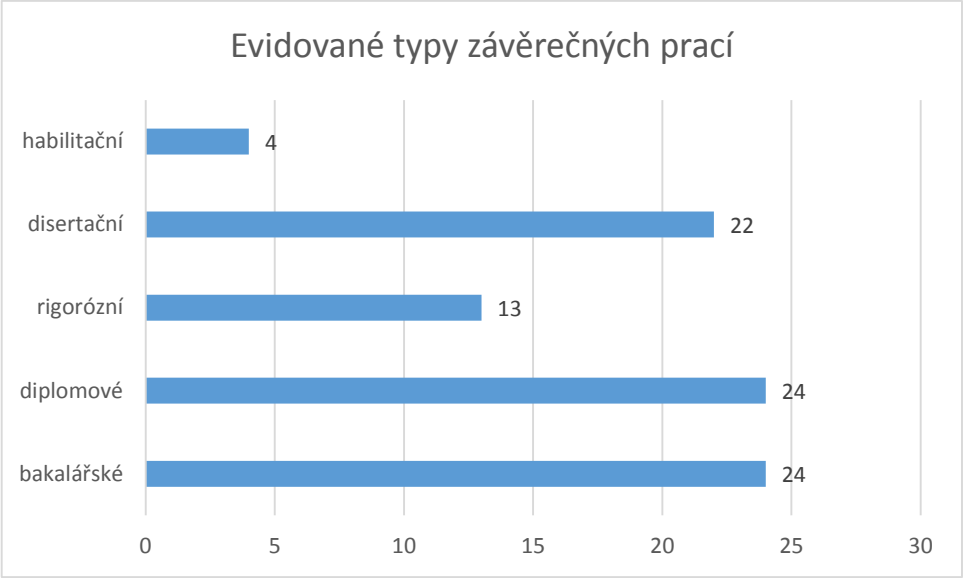


Graf 3 Odpovědné pracoviště (zdroj: autor)

Přetrvává trend centralizovaného řešení, s převahou výpočetního/informačního centra jakožto odpovědného útvaru, následovaného knihovnou. Klesající význam knihoven je pravděpodobně dán přechodem evidence eVŠKP do praxe a větším využitím informačních systémů školy pro evidenci eVŠKP (viz Graf 8).

3.4.2 Evidované typy závěrečných prací

Na otázku týkající se evidovaných typů závěrečných prací odpověděli všichni respondenti (viz Graf 4).

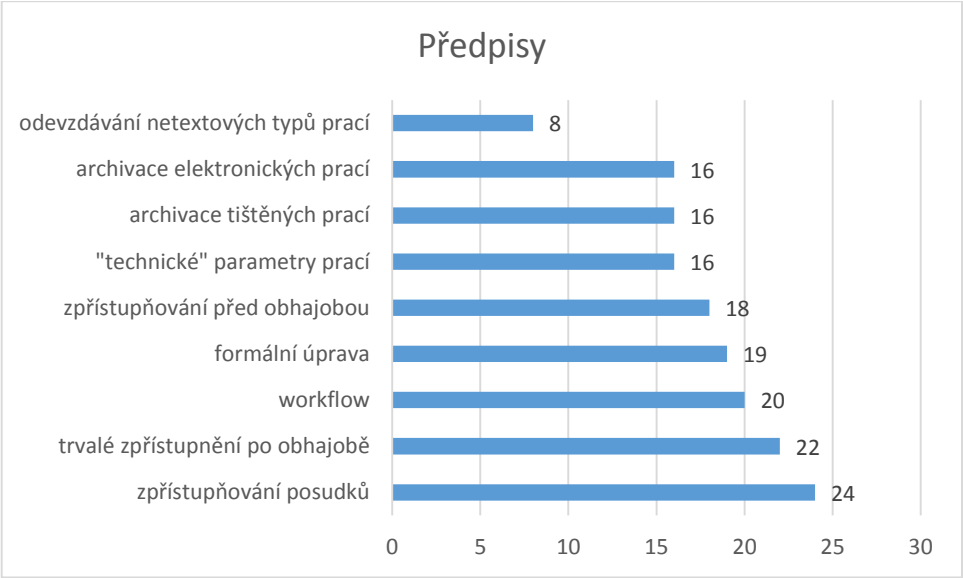


Graf 4 Evidované typy závěrečných prací (zdroj: autor)

Vysoká škola polytechnická Jihlava eviduje pouze práce bakalářské, škola má akreditovány pouze bakalářské programy. Pouze 4 školy evidují mj. i práce habilitační.

3.4.3 Předpisy regulující odevzdávání, uchovávání a zpřístupňování eVŠKP

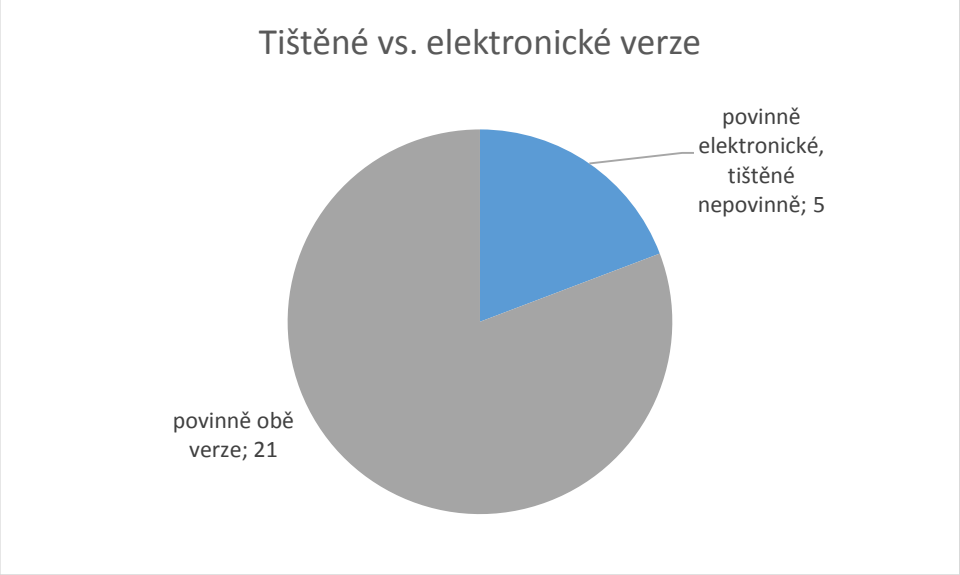
Oproti dřívějším dotazníkům byl rozšířen okruh předpisů týkajících se VŠKP, které respondenti mohli zaškrtnout v odpovědi na dotaz na platné předpisy na škole nebo fakultě (viz Graf 5).



Graf 5 Předpisy (zdroj: autor)

Celkem 18 škol v předpisech stanovuje zpřístupňování eVŠKP před obhajobou, jak vyžaduje § 47b vysokoškolského zákona (5), 22 škol upravuje trvalé zpřístupnění prací po obhajobě a 23 škol trvalé zpřístupňování posudků. Oproti dřívějším dotazníkům je tedy v předpisech vidět výrazná formalizace workflow a zpřístupňování prací. Také formální úprava prací je řešena již velmi často ($^{3}/_{4}$ škol oproti polovině v roce 2009). Nejméně je řešena problematika archivování tištěných a elektronických prací, která je zmíněna v předpisech 62 % VVŠ.

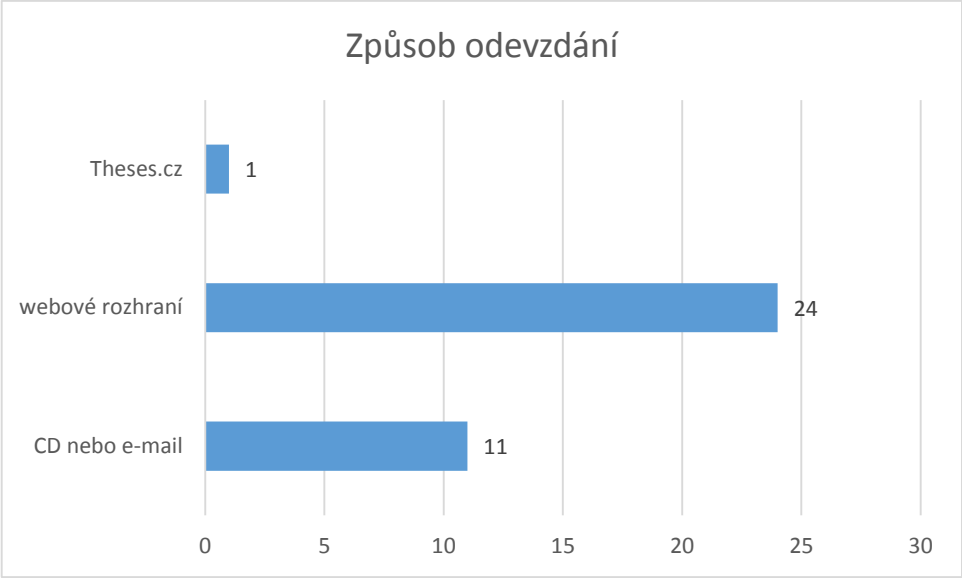
3.4.4 Odevzdávání VŠKP



Graf 6 Tištěné vs. elektronické verze (zdroj: autor)

Samostatnou otázkou bylo zkoumáno, jaké verze prací jsou odevzdávány povinně (viz Graf 6). Nejčastěji, a to na 21 školách, je povinné odevzdat tištěné i elektronické verze zároveň. Pět škol požaduje práce pouze v elektronické podobě, tištěné práce jsou vybírány nepovinně. Pozitivním zjištěním je, že všechny veřejné vysoké školy v ČR elektronickou verzi požadují odevzdat povinně.

Způsob vybírání eVŠKP zachycuje Graf 7. Respondenti mohli uvést více možných způsobů, pokud se např. na fakultách liší.

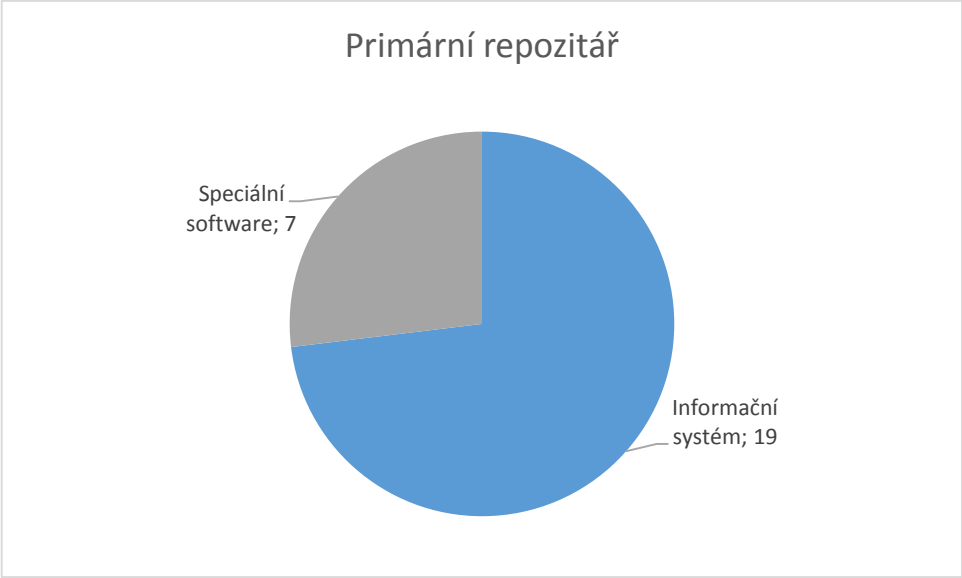


Graf 7 Způsob odevzdání (zdroj: autor)

Kromě dvou škol jsou elektronické verze VŠKP odevzdávány nahráním přes webové rozhraní informačního systému školy. Jedenáct škol podporuje (většinou alternativně k webovému rozhraní např. u netextových příloh nebo např. jen na jedné fakultě) odevzdávání eVŠKP na CD-ROM nebo zaslání práce e-mailem (tuto možnost v roce 2007 uvedlo 8 respondentů). Vložení práce do repozitáře pak zajišťují pověření pracovníci.

Fakulta jedné veřejné vysoké škole nechává studenty nahrávat elektronické verze prací přímo webovým rozhraním systému Theses.cz, který využívá pro sběr, evidenci prací a odhalování plagiátů. Práce samotné tato fakulta veřejnosti nezpřístupňuje.

Netextové typy prací je možné odevzdávat na 11 univerzitách (v roce 2007 umožňovalo 12 škol), z toho 8 škol má odevzdávání netextových typů prací řešeno v předpisech viz Graf 5 (3 respondenti v roce 2009).



Graf 8 Primární repozitář (zdroj: autor)

Na vysokých školách jsou již používány informační systémy, příp. v průběhu samotného průzkumu na některých školách právě probíhala implementace[14] informačních systémů, které nativně podporují sběr, evidenci a případné zpřístupnění eVŠKP. Jedná se např. o informační systémy IS MUNI (56), UIS (57) nebo IS/STAG (58). Informační systém jako primární repozitář používá 19 veřejných vysokých škol, speciální software má pouze 7 škol (v roce 2009 speciální software využívala polovina respondentů). V případě speciálního software se nejčastěji jedná o aplikaci DSpace.

Jako primární repozitář pro uložení eVŠKP se na veřejných vysokých školách nepoužívají knihovní katalogy ani systém Theses.cz[15].

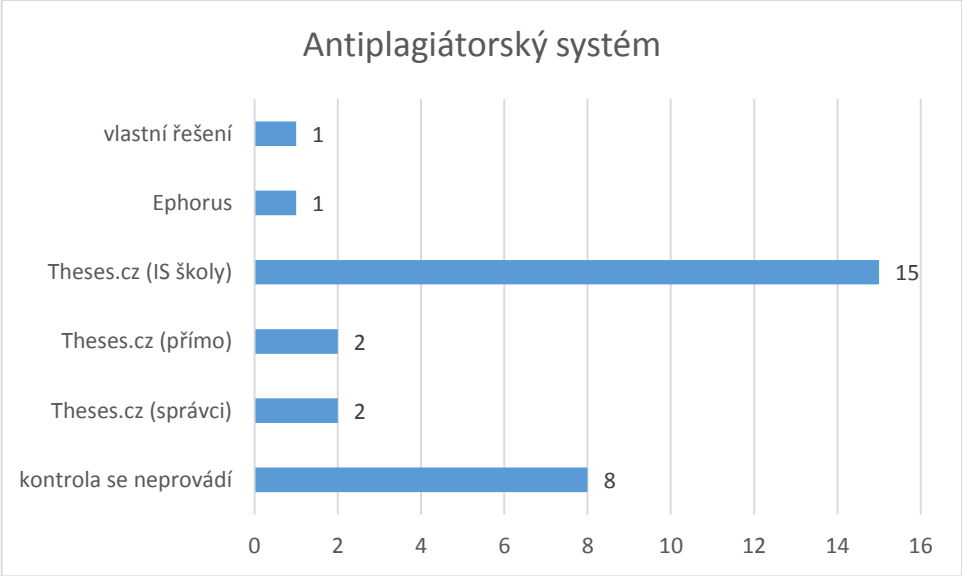
Pokračuje tak trend z minulých let, kdy práce jsou čím dál častěji odevzdávány přímo do repozitářů školy, naopak elektronické verze prací již nejsou dostupné pouze na elektronických nosičích v knihovnách.

[14] Pokud v průběhu sběru dat teprve probíhala implementace nového informačního systému, v dotazníku byl zanesen aktuální stav pro rok 2014 před spuštěním nového informačního systému.

[15] Systém Theses.cz mohou používat jako primární repozitář např. soukromé školy, které nemají tyto služby ve svém IS. Soukromé školy však nebyly předmětem tohoto výzkumu.

3.4.5 Problematika plagiátorství

Nově byla v dotazníku věnována pozornost kontrole prací na projevy plagiátorství, využití antiplagiátorských nástrojů. Školy mohly zvolit kombinaci kontrol (viz Graf 9), pokud se např. přístup lišil po fakultách.



Graf 9 Antiplagiátorský systém (zdroj: autor)

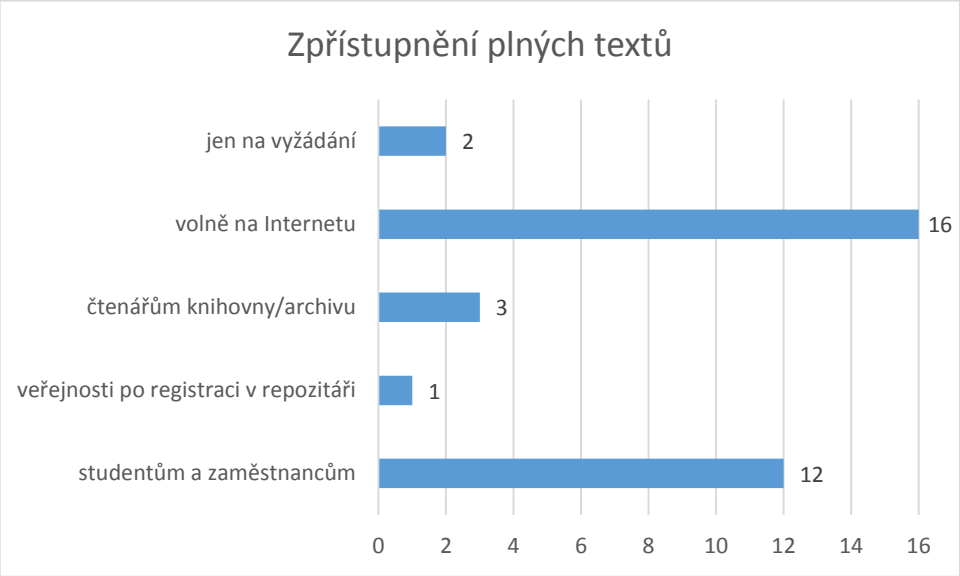
Z dotazníku vyplývá, že 18 škol aktivně využívá pro kontrolu prací aplikace Theses.cz, z toho ČVUT využívá na jedné z fakult systém Theses.cz v napojení na interní aplikaci (automatizovaný export plných textů z aplikace s ručním přenosem výsledků zpět do aplikace) a na jednom ústavu práce nahrávají studenti přímo do Theses.cz.

Pouze dvě univerzity zpřístupňují výsledky kontroly vyučujícím přímo v Theses.cz, na dvou univerzitách zpřístupňují výsledky pouze vybraným správcům za školu, v ostatních případech mají zobrazování výsledků školy integrováno v informačním systému školy.

Vysoké učení technické v Brně provozuje vlastní antiplagiátorské řešení. Systém Ephorus je využíván jako doplněk Theses.cz na Fakultě financí a účetnictví a na Fakultě podnikohospodářské Vysoké školy ekonomické v Praze, která pro všechny práce standardně využívá kontroly Theses.cz. Kontrola za pomoci aplikace na odhalování plagiátů se neprovádí na sedmi školách a částečně na ČVUT.

3.4.6 Zpřístupňování eVŠKP

Stejně jako v minulých letech byl zkoumán stav zpřístupňování, příp. nezveřejňování eVŠKP. Vysoké školy mají povinnost, danou § 47b Zákona o vysokých školách, zpřístupňovat VŠKP, posudky a průběh obhajoby prostřednictvím své databáze. Autor v dotazníku také zjišťoval, jaké jsou možnosti a důvody pro utajování prací a příloh. Veřejné vysoké školy mohly vybrat více variant aplikovaného způsobu zpřístupnění plných textů (viz Graf 10).



Graf 10 Zpřístupnění plných textů (zdroj: autor)

16 škol práce zpřístupňuje na Internetu volně. ČZU požaduje před zpřístupněním prací registraci uživatele podléhající schválení (které v případě testu repozitáře autorem proběhlo obratem, stačilo zaslat na e-mail helpdesk@czu.cz žádost se jménem, institucí a požadovaným rozmezím přístupu). V roce 2009 plné zpřístupnění eVŠKP na Internetu uvedla v dotazníku pouze VŠE v Praze (menší počet škol zveřejňujících práce bude dán menším počtem respondentů), v roce 2007 se povinně práce zpřístupňovaly pouze na 4 školách, a to nikoliv v plném „kvalifikačním“ rozsahu.

Podle odpovědí v roce 2014 zpřístupňuje práce studentům a vyučujícím celkem 12 škol, z toho:

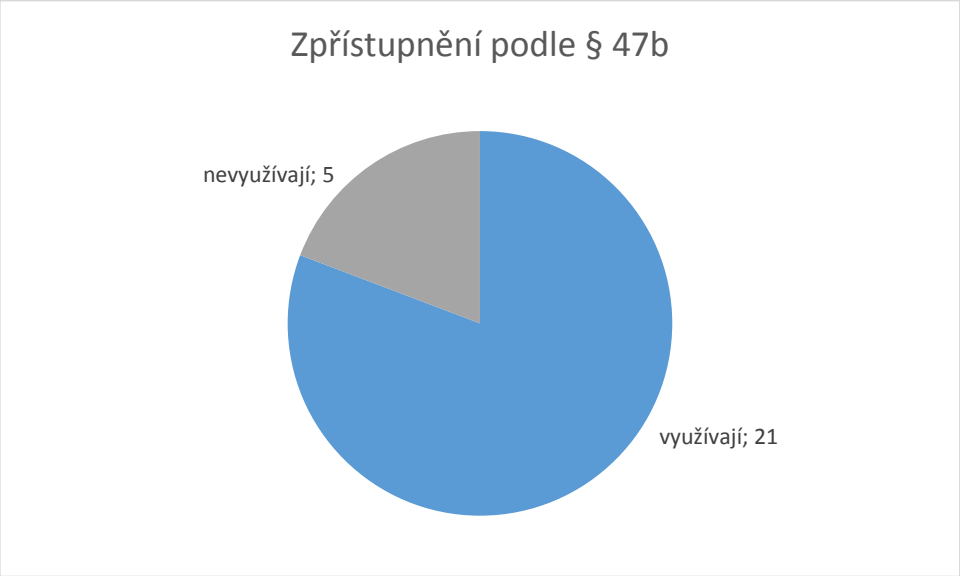
- 1 škola práce navíc zpřístupňuje externistům po registraci v repozitáři,
- 2 školy práce navíc zpřístupňují registrovaným čtenářům knihovny,

- 4 školy práce navíc zpřístupňují volně na Internetu[16],
- pro 5 škol je zpřístupnění pouze akademické obci jediný způsob a práce tak nelze považovat za volně přístupné.

Dvě školy stále uplatňují nejrestriktivnější variantu přístupu veřejnosti i studentů, a to jen na vyžádání, přičemž žádost může být zamítnuta.

Variantu zpřístupnění prací bez registrace v intranetu nebo zpřístupnění uživatelům Theses.cz neuvedla ani jedna škola.

Po sečtení všech variant můžeme konstatovat, že existuje 7 škol neumožňujících volný přístup veřejnosti k plným textům eVŠKP (tj. zpřístupňují jen akademické obci nebo jen na vyžádání), 2 školy požadují po veřejnosti osobní návštěvu knihovny nebo archivu. Zbylých 17 veřejných vysokých škol veřejnosti umožňuje plnohodnotný vzdálený přístup (z toho v jednom případě podmíněno schválením registrace e-mailem), což je významný nárůst oproti předchozím průzkumům.



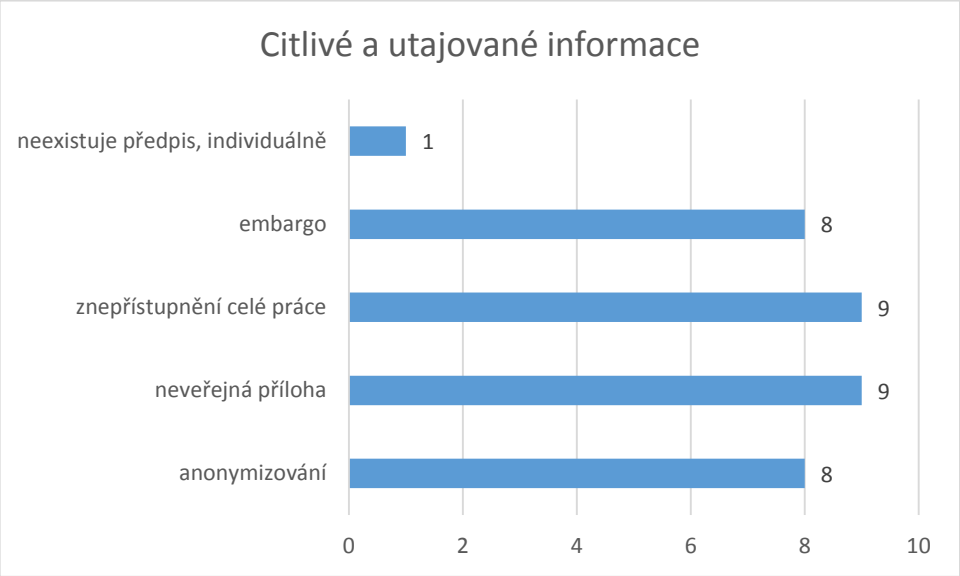
Graf 11 Zpřístupnění podle § 47b (zdroj: autor)

[16] Pokud je práce volně dostupná na Internetu, je automaticky dostupná pro studenty a zaměstnance. Respektujeme zde však varianty tak, jak je uvedly jednotlivé školy, proto v Grafu 10 počet škol zpřístupňujících práce studentům a zaměstnancům je nižší než počet škol práce zpřístupňující volně.

Zpřístupnění prací podle § 47b Zákona o vysokých školách využívá 21 škol (viz Graf 11). Pět respondentů uvedlo, že práce nezpřístupňují. Variantu zpřístupnění na základě udělení licence autorem (Creative Commons, příp. proprietární) nebo na základě zákonné licence v Autorském zákonu nevybrala ani jedna škola[17].

Oproti předchozím průzkumům tak jasně převládá výklad Zákona o vysokých školách, podle kterého autor dává odevzdáním práce souhlas se zveřejněním a nejedná se tak o porušení Autorského zákona (viz oddíl 2.1.3 *Zveřejňování a sdělování eVŠKP veřejnosti*). Stále však nejméně 5 veřejných vysokých škol práce zpřístupňuje pouze na pracovištích univerzity, další dvě školy práce teoreticky zpřístupňují na dálku pouze „na vyžádání“.

Další otázka zkoumala, za jakých okolností je možné znepřístupnit práci nebo přílohu v případě, že obsahuje citlivé nebo utajované informace. Respondenti mohli vybrat více variant, které jsou povoleny (viz Graf 12).

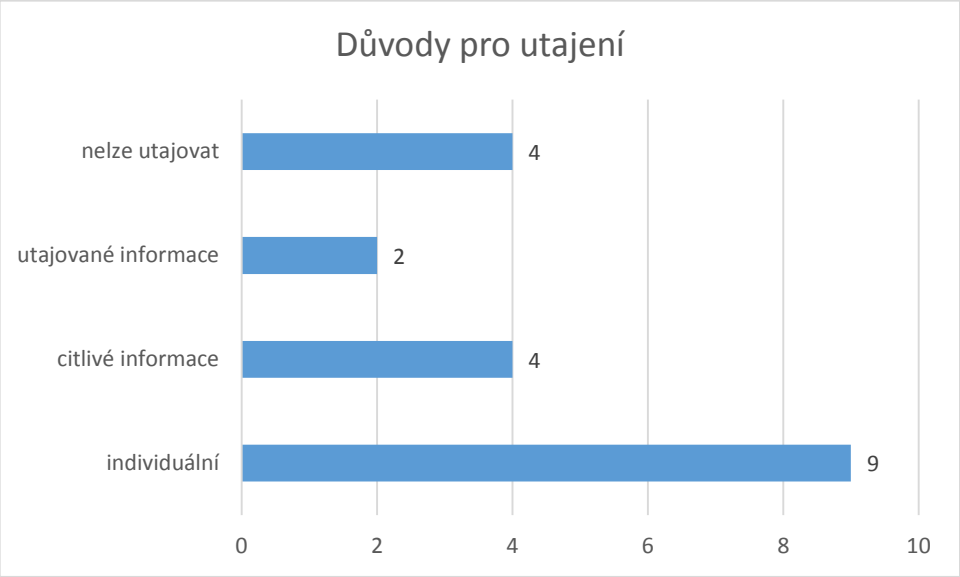


Graf 12 Citlivé a utajované informace (zdroj: autor)

[17] Pokud univerzita licence po studentech podle vnitřních předpisů požaduje, ale práce v repozitáři volně nezpřístupňuje, respondenti volili odpověď „plné texty eVŠKP nejsou dostupné veřejnosti“. Nulový počet odpovědí pro licence tedy neznamená, že by je některé školy stále nepožadovaly.

V případě uvedení citlivých, osobních nebo utajovaných informací v eVŠKP, školy – bez preferencí některé z variant – umožňují dočasné odložení zveřejnění, znepřístupnění celé práce nebo pouze přílohy s citlivými údaji. Jedna škola řeší tyto případy individuálně, 8 škol uvedlo povinnost takovéto údaje anonymizovat a práce zveřejnit celé[18]. Osm škol uvedlo kombinaci dvou až tří povolených variant řešení.

Konkrétní důvody opravňující k nezpřístupnění práce vyplňovali respondenti volným textem. Graf 13 zobrazuje odpovědi kategorizované do 4 skupin, kdy odpověď za školu mohla být započítána ve více skupinách.

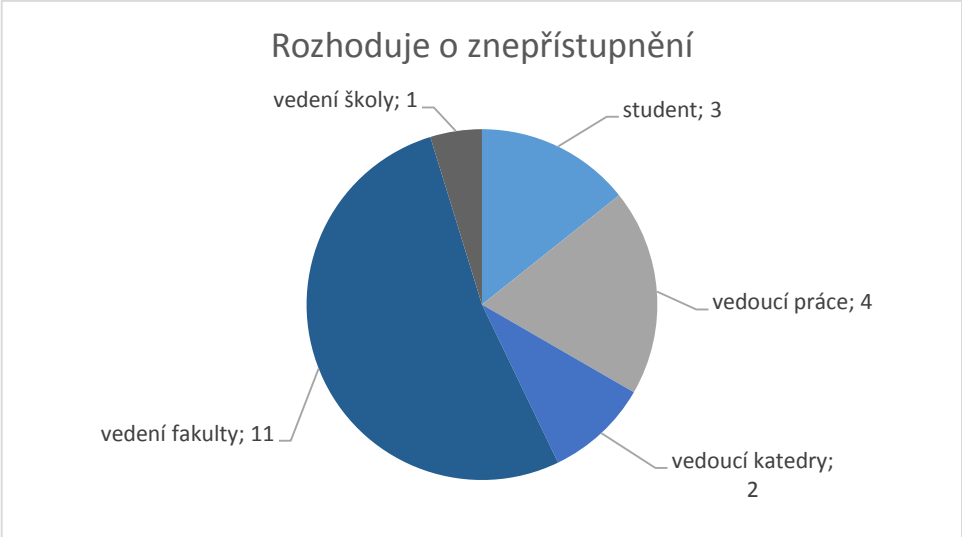


Graf 13 Důvody pro utajení (zdroj: autor)

Mezi důvody pro znepřístupnění práce nebo přílohy patří uvedení utajovaných informací (výzkum, utajované podle zákona apod.) nebo citlivých informací (osobní údaje, interní firemní informace). Nejčastěji je utajení práce řešeno individuálně, ve 4 případech naopak není povoleno práci utajit v žádném případě.

Graf 14 zachycuje, kdo na škole rozhoduje o znepřístupnění práce. Respondenti mohli vybrat právě jednu z nabízených variant odpovědí.

[18] Tato varianta asi nejvíce odpovídá smyslu § 47b Zákona o vysokých školách, který nepřipouští jakékoliv utajení práce.

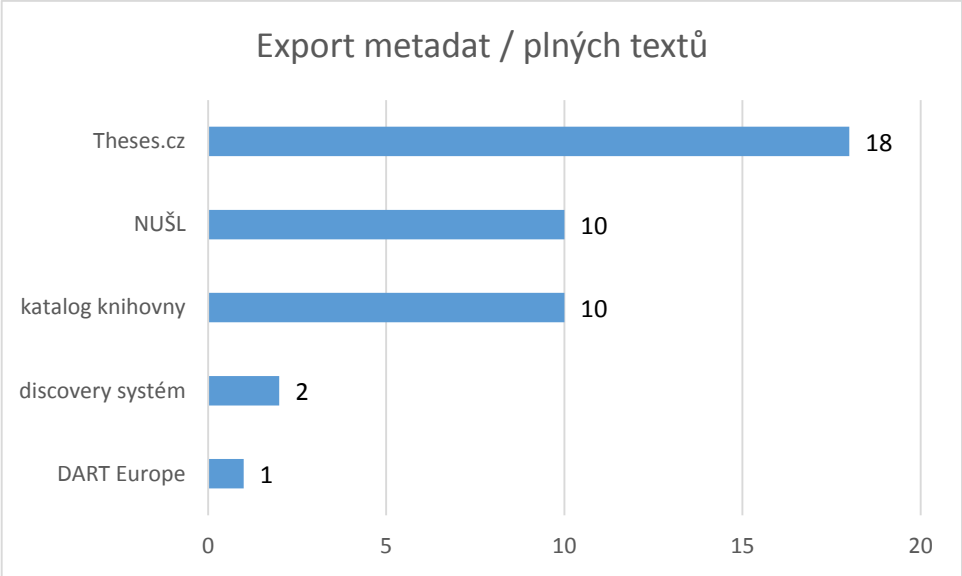


Graf 14 Rozhoduje o utajení (zdroj: autor)

O znepřístupnění své práce rozhoduje sám student pouze ve 3 případech. Nejčastěji je rozhodnutí na vedení fakulty (děkan, proděkan), méně často na vedoucím práce, vedení katedry nebo školy. Pět respondentů odpovědnou osobu neuvedlo (nepovolují utajení, příp. řešeno individuálně).

3.4.7 Exporty metadat a plných textů

Poslední okruh dotazů byl podobně jako v uplynulých letech zaměřen na export metadat a plných textů z primárního repozitáře do repozitářů externích (viz Graf 15) a používaná metadata (viz Graf 16).



Graf 15 Export metadat / plných textů (zdroj: autor)

Alespoň jednu možnost exportu zvolilo 24 z 26 škol, ale při verifikaci odpovědí jsme zjistili, že ne všechny odpovědi byly správné.

Počet škol exportujících data[19] do národního registru VŠKP – Theses.cz si žádá detailnější prozkoumání. Podle vyplněného dotazníku se jedná o 18 škol. Údaje autor disertační práce ověřoval přímo se správci Theses.cz, podle kterých do Theses.cz data posílá 21 z 26 VVŠ, z toho dvě školy přibyly v roce 2014. Metadata v Theses.cz zveřejňuje jen 17 VVŠ, které lze dohledat v rozšířeném vyhledávání Theses.cz, ostatní školy využívají Theses.cz jen pro podporu vyhledávání duplicit. Celkem do Theses.cz posílá data 43 škol se započítáním i škol soukromých, středních a zahraničních.

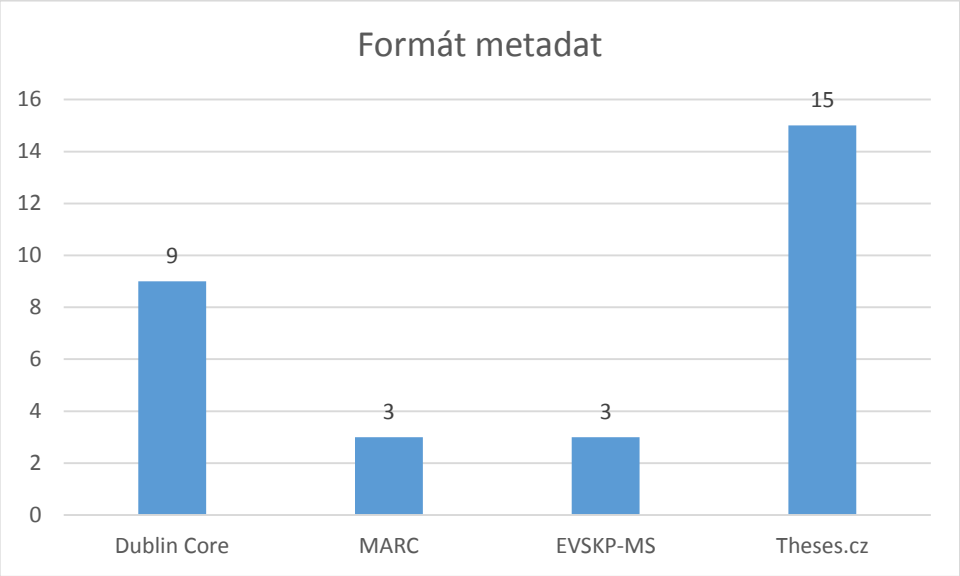
Pro připomenutí, registr Theses.cz vznikl v roce 2008 jako projekt 17 veřejných vysokých škol. Do průzkumu v roce 2009 se zapojilo 10 účastníků projektu, z nichž 2 školy do doby vyplnění dotazníku ještě neexportovaly do Theses.cz metadata.

[19] Jedná se o export metadat obsahujících odkaz na plný text. Zda je následně do externího registru stahován i plný text záleží na povaze registru, na politice zveřejňování plných textů v primárním repozitáři a příp. interní dohodě mezi školou a správcem externího registru.

Dále podle dotazníku deset univerzit exportuje data do knihovního systému a deset univerzit do Národního úložiště šedé literatury (kvalifikační práce obhajované v letech 2013–2014 jsou v NUŠL indexovány z 12 VVŠ). Dvě školy v rámci možné volné odpovědi uvedly export metadat do školou používaného discovery systému[20].

Vysoká škola ekonomická v Praze je jedinou univerzitou v ČR, která pravidelně exportuje metadata disertačních prací do evropského registru DART-Europe E-theses Portal (http://www.dart-europe.eu).

Žádná z univerzit neuvedla, že by jejich eVŠKP byly indexovány v systému OpenAIRE (http://openaire.eu), pravděpodobně z důvodu prozatím malé propagace této skutečnosti Národní technickou knihovnou – provozovatelem NUŠL. V registru OpenAIRE lze totiž nalézt i metadata eVŠKP škol zapojených v NUŠL, díky sklízení metadat přímo ze systému NUŠL.



Graf 16 Formát metadat (zdroj: autor)

Školy mohly volit více formátů metadat, které podporují (viz Graf 16). 15 škol pro export používá metadata v proprietárním formátu Theses.cz, pouze 3 školy používají standard EVSKP-MS (1) připravený Komisí eVŠKP. Vyšší zastoupení formátu Theses.cz je dáno

[20] Volba exportu metadat do discovery systému by určitě byla vhodná jako samostatná varianta v navazujícím průzkumu. Lze předpokládat, že metadata eVŠKP jsou indexována ve většině discovery systémů používaných na školách v ČR.

pravděpodobně zapojením většiny škol do systému Theses.cz, který svůj vlastní formát primárně doporučuje. Nižší počet odpovědí u volby EVSKP-MS může být dán neznalostí všech možností systému IS/STAG (viz níže). Pouhých 9 škol uvedlo podporu standardu Dublin Core, jen 3 školy využívají doplňkově export metadat na bázi MARC. Trend preference formátu Theses.cz pro export do Theses.cz je jednoznačný.

Graf 17 Způsob exportu (zdroj: autor)

Všech 23 respondentů, kteří odpověděli na dotaz na způsob exportování/předávání metadat eVŠKP mimo školu, údajně používají standard OAI-PMH, v jednom případě doplněný proprietárním řešením ad hoc exportu metadat. Při verifikaci údajů autor narazil na rozpor s tím, jakým způsobem se data odevzdávají do Theses.cz v praxi – protokol OAI-PMH pro export metadat do Theses.cz používá jen 6 VVŠ[21], ostatní školy používají proprietární způsob Theses.cz využívající pro zasílání XML souborů metodu POST protokolu HTTP.

Zarážející je nízká podpora standardu Dublin Core (9 odpovědí viz výše) v případě, kdy 23 univerzit by mělo podporovat harvestování protokolem OAI-PMH, pro který je Dublin Core povinným metadatovým formátem (31). Při evaluaci odpovědí autor výzkumu experimentálně ověřil, že školy využívající IS/STAG mohou využít jak formát Theses.cz, tak i EVSKP-MS (oba formáty jsou obsahem prakticky rovnocenné). Překvapující je však

[21] Jedná se převážně o školy používající repozitáře na bázi IS MUNI nebo DSpace.

absence formátu Dublin Core v implementaci OAI-PMH serveru IS/STAG, jedná se tak nepochybně o chybu v implementaci jinak velmi povedeného OAI-PMH serveru podporujícího oba formáty pro popis metadat eVŠKP.

Při implementaci podpory metadatového formátu nekvalifikovaný Dublin Core v OAI-PMH lze vycházet z doporučeného mapování, zpracovaného v kapitole 4 této disertační práce.

3.5 Závěr kapitoly

V závěru kapitoly shrneme hlavní poznatky z provedeného dotazníkového šetření a odpovíme na otázky stanovené na začátku.

Poprvé se podařilo v takto komplexní podobě zmapovat stav zpřístupňování eVŠKP na všech veřejných vysokých školách v ČR. Obdobný průzkum provedený Odbornou komisí pro otázky elektronického zpřístupňování vysokoškolských kvalifikačních prací AKVŠ ČR se konal v roce 2009, kdy na dotazník odpovědělo o deset škol méně.

Průzkum ukázal, že oproti předchozím rokům již všechny veřejné vysoké školy sbírají eVŠKP v elektronické podobě. Povinné odevzdání elektronické verze je u většiny škol doplněno i o povinné odevzdávání verze tištěné. V souvislosti s digitalizací spisové služby na školách lze předpokládat v následujících letech větší příklon k odevzdávání pouze verze elektronické.

Problematika zpřístupňování eVŠKP je v současnosti nejčastěji v odpovědnosti výpočetních či informačních center, méně často v odpovědnosti knihoven. Menší počet knihoven jakožto odpovědných útvarů pravděpodobně souvisí se začleněním workflow odevzdávání a zpřístupňování eVŠKP mezi rutinní činnosti školy, podporované informačním systémem.

Ve většině škol studenti vyplňují metadata a nahrávají plné texty přes webové rozhraní přímo do informačního systému školy, pouze 7 škol používá pro evidenci eVŠKP specializovaný software, nejčastěji DSpace.

Velmi pozitivním trendem oproti předchozím průzkumům je razantní nárůst zpřístupňování eVŠKP volně na Internetu bez registrace, což činí 16 VVŠ. Ostatní školy veřejnosti práce nezpřístupňují buď v žádné podobě, nebo uplatňují různé bariéry přístupu (nutnost zažádat o zpřístupnění práce k osobnímu nahlédnutí v knihovně nebo archivu, nutnost vyžádat si přístupové údaje, získat souhlas ke zpřístupnění apod.).

Nejčastěji jsou práce zveřejněny na základě souhlasu autora uděleného odevzdáním práce do databáze školy, v souladu s § 47b Zákona o vysokých školách. Oproti předchozím rokům je znatelný ústup od podepisování licenčních smluv, v praxi se nepotvrdila potřeba licencí v písemné podobě.

Osm škol neumožňuje v žádné formě utajení práce nebo přílohy, jejich studenti musí citlivá a utajovaná data anonymizovat a práci zveřejnit celou. Tento způsob plně odpovídá požadavkům § 47b Zákona o vysokých školách, který nepřipouští jakékoliv utajování plných textů nebo výsledků obhajoby. Ostatní školy umožňují, nejčastěji po individuálním posouzení důvodů, utajení práce, přílohy nebo dočasné odložení zveřejnění, přičemž ani jedna z variant není výrazněji preferována. Pokud dříve školy neměly jasno, jak nakládat s citlivými a utajovanými informacemi, v současné době již podle odpovědí můžeme usuzovat, že každá škola si interně stanovila postupy, jak s takovýmito informacemi zacházet, i když to možná plně neodpovídá požadavkům zákona.

Většina veřejných vysokých škol využívá služeb aplikací na podporu odhalování plagiátů. 21 veřejných vysokých škol využívá systém Theses.cz, z toho 15 má podle odpovědí v dotazníku výsledky kontroly integrovány přímo do informačního systému školy.

Větší využití discovery služeb pro vyhledávání eVŠKP a propagace metadat eVŠKP do dalších repozitářů by umožnily snazší nalezení zajímavých vysokoškolských kvalifikačních prací, které by jinak v repozitářích univerzit zůstaly nepovšimnuty. Vzhledem k aktuální poptávce knihoven v ČR po discovery službách, indexujících i metadata eVŠKP, autor v kapitole 6 uvádí doporučení pro výběr těchto služeb.

I nadále přetrvává preference formátu Theses.cz pro popis metadat eVŠKP. Nově je již na školách více podporován standardizovaný protokol OAI-PMH, který je ve světě doporučován pro automatizované, pravidelné harvestování metadat z lokálních repozitářů. Implementace OAI-PMH serveru v repozitáři univerzity umožňuje snadné šíření metadat i do dalších repozitářů v ČR a ve světě. V optimálním případě by u OAI-PMH měla být data popsána nejen obecným metadatovým standardem Dublin Core, ale i dalšími formáty umožňujícími detailnější popis eVŠKP (např. EVSKP-MS, Theses.cz, MARCXML). Z výše uvedených důvodů lze školám využívajícím IS/STAG doporučit, aby v rámci OAI-PMH serveru řádně podporovaly i standard Dublin Core a více napomáhaly rozšiřování metadat svých prací přes tento protokol.

Univerzitám lze doporučit spolupráci nejen s projektem Theses.cz (zajišťujícím v ČR jednotné rozhraní pro vyhledávání eVŠKP a kontrolu prací na plagiátorství), ale i s Národním úložištěm šedé literatury, díky kterému se metadata zároveň dostanou i do indexu OpenAIRE. V případě podpory otevřeného přístupu k disertačním pracím v repozitáři školy autor doporučuje spolupráci i s evropským repozitářem DART-Europe E-theses Portal (viz oddíly 2.3.1 a 4.3.3).

4 Mapování metadat eVŠKP

Autor práce se dlouhodobě zabývá problematikou metadatového popisu vysokoškolských kvalifikačních prací v digitálních repozitářích. V rámci Komise eVŠKP s Evou Bratkovou a ve spolupráci s dalšími členy komise na základě analýzy existujících mezinárodních metadatových standardů vypracovali *Standardizační soubor metadatových prvků určených pro popis vysokoškolských kvalifikačních prací obhajovaných na vysokých školách v ČR a pro přenos souborů EVSKP-MS* (1) a související standardy pro popis fyzických osob PersCZ (2) a korporací CorpCZ (3).

Analýzou zahraničních metadatových standardů pro popis a komunikaci vysokoškolských kvalifikačních prací, představením návrhů a reálné verze standardů EVSKP-MS, PersCZ a CorpCZ se důkladně věnuje v 5. kapitole své disertační práci spoluautorka zmiňovaných dokumentů Eva Bratková (59). Uvedené metadatové standardy jsou výsledkem výzkumné a vývojové činnosti, v případě autora této disertační práce v rámci centralizovaného rozvojového projektu MŠMT C1/2008 *Národní registr VŠKP a úložiště závěrečných prací se službou na odhalování plagiátů*. Přínos autora při tvorbě standardu EVSKP-MS je především v oblasti technických a administrativních metadatových prvků, ve formalizaci zápisu metadat do XML včetně validace metadat a ve vypracování doporučení pro využití protokolu OAI-PMH při komunikaci metadat (případová implementace OAI-PMH serveru viz podkapitola 4.3).

Autor navazuje na již spoluautorkou popsané analýzy a výsledky ve výše zmíněné disertační práci. V následující analýze zkoumá možnosti mapování standardu EVSKP-MS na další standardy, doporučení experimentálně ověřil nad daty Databáze kvalifikačních prací VŠE (popis viz oddíl 2.3.2 a podkapitola 4.3) a při exportu metadat do mezinárodního repozitáře DART-Europe E-thesis Portal (viz oddíl 2.3.1).

4.1 Metadatové standardy pro popis eVŠKP

Porovnání metadat ETD-MS a dalších zahraničních metadatových standardů pro popis závěrečných kvalifikačních prací provedl Neil Godfrey v dokumentu *Electronic Theses and Dissertation Metadata Schema (ETD-MS) for Australia?* (60). Výsledkem analýzy je tabulka publikovaná v článku Neila Godfreye, porovnávající jednotlivé prvky mezi metadaty Eprints,

DSpace, kvalifikovaný Dublin Core, UKETD-DC, XMetaDiss, TEF, MARC, ETD-MS a MODS.

Cílem níže uvedené analýzy je provést obdobné mapování pro český metadatový standard EVSKP-MS.

Prvky standardu EVSKP-MS je možné zakódovat do mnoha dalších metadatových sad. Z výsledků provedeného průzkumu (viz kapitola 3 *Průzkum zpřístupňování vysokoškolských kvalifikačních prací v roce 2014*) víme, že v ČR se využívají především metadatové sady Theses.cz, EVSKP-CZ a Dublin Core, příp. MARC 21 v katalozích knihoven.

Metadatový standard *Dublin Core Metadata Element Set* (*ISO 15836:2003*) (61) definuje základní elementy pro popis zdrojů, přičemž popisovaný zdroj ani implementační detaily nejsou ve standardu specificky definovány. Je základem doporučení pro import metadat např. v *DRIVER Guidelines* (62), *OpenAIRE Guidelines: For Literature repositories* (63) a repozitáře DART-Europe E-theses Portal (viz oddíl 2.3.2). Standardy Dublin Core Metadata Iniciative (zkráceně DCMI) – *Dublic Core Metadata Element Set* (zkráceně DC Set) a *DCMI Metadata Terms* (64) (zkráceně DCMI Terms) – jsou pro svoji obecnou použitelnost základem mnoha dalších metadatových standardů a aplikací. Standard DC Set je povinné podporovat v rámci OAI-PMH serverů, které repozitáře OpenAIRE, DRIVER a DART-Europe E-theses Portal využívají pro harvestování metadat.

Metadatová sada prvků Theses.cz (65) (zkráceně PTS[22]) byla připravena týmem MUNI při implementaci národního repozitáře VŠKP Theses.cz. Metadatové formáty EVSKP-MS a PTS jsou podporovány na importu metadat do Theses.cz. Obsahově PTS vychází ze standardů EVSKP-MS, PersCZ a CorpCZ, kdy uvedené formáty mixuje, některé prvky nad rámec Dublin Core definuje samostatně ve vlastním jmenném prostorem s prefixem pts.

Oproti formátům EVSKP-MS, PersCZ a CorpCZ je v PTS používáno kvalifikátorů, tj. tečkové notace namísto základních metadatových prvků doplněných o atributy (např. <pts:title.translated> namísto <dc:title evskp:typeTranslated="translated">).

[22] Abychom v kapitole formálně odlišili metadatový formát Theses.cz od názvu národního repozitáře VŠKP Theses.cz, budeme používat pro označení metadatového formátu Theses.cz zkratku PTS, použitou jeho tvůrci jako prefix proprietárních prvků tohoto formátu.

Kvalifikátory v XML nemusí být plně podporovány při importu metadat vyhledávači a může tak dojít k sémantické ztrátě informace při importu dat ve formátu PTS.

Standardy *Metadata Encoding and Trasmission Standard* (METS) a *Metadata Object Description Schema* (MODS) spravuje Kongresová knihovna. Lze je využít v oblasti administrativních, právních a strukturálních metadat (METS) a v oblasti popisných metadat (MODS). Pravděpodobně pro svoji komplikovanost však nejsou v praxi na našich univerzitách pro popis eVŠKP využívány.

Standard *ETD-MS v1.1: an Interoperability Metadata Standard for Electronic Theses and Dissertations* (zkráceně ETD-MS) (30) spravovaný mezinárodní organizací NDLTD (viz oddíl 2.3.1) byl hlavním formátem, ze kterého autoři EVSKP-MS při přípravě standardu vycházeli. Pokud to bylo možné, přejímali definice metadatových prvků z ETD-MS, ať už se jednalo o prvky DC Set (jmenný prostor označen ve standardu EVSKP-MS prefixem dc) nebo o prvky specifické pro ETD-MS (jmenný prostor označen ve standardu EVSKP-MS prefixem theses). V případě nutnosti, převážně u technických a administrativních prvků, byly přidány navíc prvky vlastní (jmenný prostor označen ve standardu EVSKP-MS prefixem evskp), příp. prvky rozšířeného Dublin Core – DCMI Terms (jmenný prostor označen ve standardu EVSKP-MS prefixem dcterms).

Standard ETD-MS obsahuje doporučené mapování prvků ETD-MS do formátu MARC 21, jehož cílem je sjednotit praxi knihoven preferujících některý ze skupiny MARC formátů. Katalogizátoři na univerzitách v ČR již bezesporu mají zkušenost s popisem tištěných VŠKP v katalozích vlastních knihoven, mapování do MARC 21 proto nebude předmětem této analýzy.

4.2 Mapování prvků formátu EVSKP-MS

Rozšíření nabídky podporovaných metadatových formátů při exportu z repozitářů umožní snazší, širší propagaci eVŠKP do dalších repozitářů. Na základě provedených analýz, mj. vzhledem k objevené chybě v implementaci OAI-PMH v IS/STAG (viz oddíl 3.4.7), se ukazuje především potřeba vypracování mapování prvků standardu EVSKP-MS (1) na prvky formátu PTS (65) vzhledem k jeho velkému zastoupení v ČR a na nekvalifikované prvky formátu DC Set (61), jehož implementace je povinná pro protokol OAI-PMH. Dále jsou v analýze zahrnuty dva další významné formáty, nejrozšířenější metadatový formát popisující

eVŠKP v zahraničí ETD-MS (30) a metadatový standard *DCMI Metadata Terms* (64) pro rozšířený popis zdrojů.

Mapování prvků formátů EVSKP-MS je využitelné mj. pro import metadat do Národního úložiště šedé literatury, podporujícího na straně importu metadat eVŠKP formát EVSKP-MS[23]. Obecné prvky DC Set jsou preferovány na straně importu např. repozitářem DART-Europe E-theses Portal (viz 2.3.1). Metadatový set DCMI Terms je velmi často využíván jako výchozí pro mnohá další metadatová schémata a ontologie (např. Schema.org, pro které bylo mapování prvků DCMI Terms na DC Set již zpracováno pracovní skupinou DCMI).

Nejvíce komplikací se ukázalo při mapování na metadatový formát PTS, který zavádí velké množství vlastních metadatových prvků a využívá upřesňujících kvalifikátorů, jejichž podpora v aplikacích pro správu metadat není povinná a prvek pak bývá interpretován jako nekvalifikovaný.

Strukturou metadat se PTS liší od EVSKP-MS především u komunikačních metadat, kdy namísto atributů využívá vnořených prvků. Pro příklad srovnejme prvky pro popis plných textů z obou standardů:

Formát PTS:

```
<pts:presentation.file>
  <pts:url>http://is.muni.cz/th/99840/diplomka.pdf</pts:url>
  <pts:ctype>thesis</pts:ctype>
  <pts:mtype>application/pdf</pts:mtype>
  <pts:size>473643</pts:size>
  <pts:author>1</pts:author>
</pts:presentation.file>

<pts:get.file>
  <pts:url>http://is.muni.cz/th/99840/diplomka.pdf</pts:url>
  <pts:ctype>thesis</pts:ctype>
  <pts:author>1</pts:author>
</pts:get.file>
```

[23] Metadatový formát NUŠL pro popis šedé literatury byl inspirován standardem EVSKP-MS. Z formátu EVSKP-MS přejímá strukturu popisu prvků, prvky obecné a údaje související s vysokoškolskou prací v prvku theses:degree. Autor disertace se jako člen vývojového týmu a projektu NUŠL mj. spolupodílel na přípravě metadatového formátu NUŠL, posudek k formátu zpracovávala Eva Bratková.

Formát EVSKP-MS:

```
<evskp:fileProperties fileID="1223" fileType="thesis" fileName="diplomka.pdf" fileSize="1145628"
  format="application/pdf">Hlavní práce</evskp:fileProperties>

<evskp:transfer accessRights="public" fileID="1249164">
http://is.muni.cz/th/99840/diplomka.pdf</evskp:transfer>
```

Formát PTS definuje zvlášť prvek pro popis prezentace souboru v lokálním repozitáři školy <pts:presentation.file> a zvlášť prvek pro pokyn Theses.cz ke stažení plného textu <pts:get.file>. U těchto XML prvků PTS využívá vnořených podprvků, obdobu EVSKP-MS atributů. Standard EVSKP-MS používá prvek popisující vlastnosti souboru <evskp:fileProperties> a prvek definující dostupnost souboru <evskp:transfer>.

Výsledkem komparační analýzy uvedených formátů je vypracované doporučení pro mapování prvků EVSKP-MS na další formáty, které je uvedeno ve formě tabulky v Příloze IV. První sloupec tabulky zobrazuje všechna metadata ze standardu EVSKP-MS (prefix evskp). Sloupce následující obsahují doporučené odpovídající metadatové prvky v metadatových formátech PTS (prefix pts), ETD-MS (prefix thesis), DC Set (prefix dc) a DCMI Terms (prefix dcterms). Prefixy u prvků označují, z jakého metadatového formátu použitý prvek pochází. Pokud cílový metadatový formát neobsahuje odpovídající volitelný prvek nebo mapování by bylo nevhodné (např. kvůli zvyšování výskytu opakovatelných prvků), zůstává odpovídající buňka tabulky nevyplněna.

Z analýzy vyplynulo, že v některých případech nelze provést mapování prvků 1:1, ale je zapotřebí transformace, např. mapování jednoho prvku EVSKP-MS na více prvků formátu PTS, s odlišením podle atributu prvku EVSKP-MS.

Vysvětlující komentáře k mapování prvků dle tabulky v Příloze IV jsou uvedeny níže, ve členěné podle standardu EVSKP-MS. Pokud je mapování v tabulce jednoznačné, není komentář k mapování daného prvku uváděn.

Název VŠKP

Prvek dc:title obsahuje volitelný atribut evskp:typeTranslated="translated" označující hlavní název v cizím jazyce. Při mapování do PTS je nutné překlady názvu uvést v samostatném prvku pts:title.translated. U ostatních standardů mapujeme 1:1 do prvku dc:title (včetně možného zachování parametru xml:lang, pokud je podporován).

Podnázev VŠKP

Při mapování do standardu ETD-MS by měl být podle definice ETD-MS použit cílový prvek dc:title.alternative. V příkladu XML kódování uvedeném ve standardu ETD-MS je však použit metadatový prvek dcterms:alternative. Jedná se tak o nekonzistenci v tomto standardu. V doporučení se autor proto přiklání k použití prvku dc:title.alternative, který standard ETD-MS explicitně uvádí v definici prvků.

Autor VŠKP

Povinný prvek dc:creator je možné ve standardu EVSKP-MS zapsat 1) jako text ve formátu „*Příjmení, Křestní jméno*“ nebo „*Příjmení, Křestní jméno; doplňující informace*“, 2) strukturovaně podle standardu PersCZ. Při mapování do PTS v prvním případě použijeme jako cílový prvek dc:creator, v druhém případě prvek pts:creator určený pro strukturovaný zápis obdobný PersCZ.

V PTS je povinné uvést buď první, nebo druhý prvek, jak vyplývá z příkladu ke standardu. Prvek pts:creator je opakovatelný, prvek dc:creator není (což odpovídá definici EVSKP-MS a definici VŠKP, které mají mít jen jednoho autora).

Výhodou použití jen jednoho prvku dc:creator u EVSKP-MS je snazší automatizovaná validace metadat na povinný prvek, univerzálnost daná použitím prvku ze standardu DC Set, přičemž text prvku je vždy strojově čitelný i v případě použití PersCZ (při ignorování specifických prvků PersCZ, se zachováním textového obsahu prvků).

Věcný popis VŠKP

Jednotlivé termíny věcného popisu VŠKP se u EVSKP-MS a PTS oddělují středníkem, v případě ETD-MS je použito opakování prvku pro každý termín zvlášť. Věcný popis VŠKP je povinný pouze ve standardu ETD-MS, pokud by zdrojová metadata věcný popis neobsahovala, bylo by nutné při mapování do ETD-MS termíny doplnit katalogizátorem. Standardy Dublin Core v definici prvku doporučují pouze specifikovat použitý kontrolovaný slovník, předpokládáme proto opakování prvku dc:subject.

Instituce archivující anebo zpřístupňující VŠKP

Standard EVSKP-MS připouští dvě formy zápisu, volným textem (přičemž pro zápis podřízené jednotky je použito formalizovaného zápisu „*Instituce. Podřízená jednotka*“), nebo strukturovaný zápis CorpCZ. Při mapování do formátu PTS je nutné název vysoké školy zapsat do prvku dc:publisher a název fakulty do proprietárního prvku pts:publisher.faculty.

Vedoucí nebo oponent VŠKP

Metadatový formát PTS jako jediný používá dva proprietární prvky, proto je zapotřebí transformace prvku EVSKP-MS podle atributu thesis:role.

Data vytvoření, odevzdání, podání či obhajoby VŠKP

Metadatové formáty PTS a ETD-MS jako povinně uváděné datum stanovují datum vytvoření, prvek dcterms:dateSubmitted (PTS) nebo dc:date (ETD-MS). Autoři EVSKP-CZ se rozhodli, vzhledem k možnosti automatizace tvorby metadat a většímu významu data obhajoby VŠKP, pro povinný prvek dcterms:dateAccepted.

Pro mapování data do PTS je proto použit jeden z následujících prvků z EVSKP-MS, pokud je evidován (v tomto pořadí preferencí):

1. dcterms:created (v EVSKP-MS Datum vytvoření VŠKP, volitelný)
2. dcterms:dateSubmitted (v EVSKP-MS Datum odevzdání či podání VŠKP, volitelný)
3. dcterms:dateAccepted (v EVSKP-MS Datum obhajoby VŠKP, povinný)

Není možné opakovat prvek Data vytvoření VŠKP v metadatových formátech PTS ani ETD-MS, prvky jsou definovány jako neopakovatelné. Při mapování do PTS je však možné mapování dcterms:dateAccepted → dcterms:dateSubmitted (Datum vytvoření v PTS) a zároveň dcterms:dateAccepted → dcterms:dateAccepted (Datum obhajoby v PTS).

Pro doplnění je nutné upozornit na riziko záměny sémantické interpretace prvků: jak již bylo uvedeno výše, formát PTS definuje jako Datum vytvoření práce prvek dcterms:dateSubmitted, přitom ve standardu DCTERMS (a potažmo v EVSKP-MS) je pro datum vytvoření určen prvek dcterms:created.

Typ VŠKP

Při mapování typu VŠKP je nutné zohlednit případný číselník podporovaný cílovým metadatovým formátem či harvestujícím serverem, příp. pokud to je možné uvést typ VŠKP ve více variantách. V případě standardu Dublin Core je vhodné v opakujícím se prvku dc:type použít klasifikace podle slovníku *OpenAIRE Guidelines* (63), tj. např. hodnotu info:eu-repo/semantics/doctoralThesis pro disertační práce.

Médium (formát souboru) VŠKP

Pouze metadatový formát PTS zavádí vlastní proprietární prvek pts:mtype. Ostatní formáty používají prvek dc:format nebo dcterms:medium.

Identifikátor VŠKP

Jednoznačný identifikátor VŠKP dc:identifier při mapování do formátu PTS uvádíme jak v prvku dc:identifier (jako odkaz do repozitáře), tak v proprietárním prvku pts:thesis.id (povinný, interní identifikátor repozitáře).

Zachování jednotného, celosvětově unikátního identifikátoru VŠKP v cílové metadatové sadě zaručuje identifikaci identických metadatových záznamů, jejich spárování a deduplikaci např. v centrálním indexu discovery služeb (viz kapitola 6 *Výběr systému centralizovaného vyhledávání*) či umožňuje výpočet agregovaných metrik užití eVŠKP (viz kapitola 5 *Metriky pro měření užití eVŠKP v online prostředí*).

Práva k využívání VŠKP

Metadatový formát PTS používá pro textový popis práv k využívání VŠKP prvek dc:rights, podobně jako další zmiňované metadatové formáty. Na rozdíl od nich však v případě použití URL adresy požaduje její uvedení do proprietárního prvku pts:rights.href. Komise eVŠKP se rozhodla pro použití pouze prvku dc:rights, neboť odlišení URL adresy od textu je jednoduše proveditelné např. regulárním výrazem.

Akademický titul nebo vědecko-pedagogická hodnost

Metadatový formát EVSKP-MS používá strukturovaného zápisu v prvku thesis:degree ze jmenného prostoru prvků ETD-MS, který nemá jednoznačnou obdobu v obecněji zaměřených standardech Dublin Core.

Metadatový formát PTS používá proprietárního prvku pts:degree, v obdobném členění jako ETD-MS a EVSKP-MS, pouze s vlastními kvalifikátory pro studijní obor pts:degree.field (v EVSKP-MS je studijní program a studijní obor oddělen lomítkem, je tak zachována kompatibilita s ETD-MS) a pro fakultu školy přidělující titul pts:degree.grantor.faculty (v EVSKP-MS je použito formalizovaného zápisu „*Instituce. Podřízená jednotka*“ nebo strukturovaného zápisu podle metadatového formátu CorpCZ). Při mapování EVSKP-MS → PTS je tedy nutná transformace obsahu.

Identifikátor poskytovatele metadat

Prvek evskp:contact obsahuje povinný atribut contactID, identifikující odesilatele metadat číslem ve tvaru předepsaném celostátní matrikou studentů, příp. Ústavem pro informace ve vzdělávání. Obsahem prvku je nepovinné textové vyjádření jména instituce, které je využitelné např. pro archivní účely či pro zahraniční instituce bez dostupného číselníku celostátní matriky studentů.

Obdobou tohoto prvku v PTS je prvek pts:sender.id, který má uveden identifikátor odesilatele (atribut contactID z EVSKP-MS) s prefixem S jako svoji hodnotu. Dochází zde tedy ke ztrátě informace o jménu instituce.

Počet souborů VŠKP

Počet všech dokumentů a příloh tvořících VŠKP z EVSKP-MS nemá v dalších metadatových formátech obdobu. Počet souborů v cílových formátech lze dopočítat podle počtu opakujících se specifických prvků, odkazujících na plný text práce.

Popis konkrétního souboru VŠKP

Prvek evskp:fileProperties a jeho atributy se ve formátu PTS mapují na prvek pts:presentation.file a v něm vnořené prvky pts:ctype, pts:mtype, pts:size a pts:filename,

obsahující informace popisující konkrétní soubor eVŠKP. Název souboru, tj. obsah prvku evskp:fileProperties, je konkrétně mapován na prvek pts:fileinfo.

Prvek pts:presentation.file je ve vypracovaném mapování autorem disertační práce upřednostňován před prvkem pts:get.file, neboť soubor eVŠKP má být veřejnosti zpřístupňován z lokálního repozitáře univerzity, nikoliv z interního úložiště webové aplikace Theses.cz. V případě zpřístupnění plného textu v Theses.cz by bylo vhodné implementovat agregaci a zpracování statistických dat nad lokálním repozitářem univerzity a repozitářem Theses.cz (viz odstavec 5.6 *Agregace a zpracování statistických dat*) za účelem měření míry užití eVŠKP uložené ve více repozitářích.

Identifikátor odkazující na soubor tvořící VŠKP nebo archiv ZIP

URL adresa souboru pro přenos platná v době zpřístupnění metadatového záznamu EVSKP-MS je ve formátu PTS mapována na prvek pts:url, vnořený v prvku pts:presentation.file (viz odstavec výše). Ve formátech Dublin Core může být použito prvku dc:relation, příp. dc:identifier.

Informace o serveru zpřístupňujícím VŠKP

Informace o serveru zpřístupňujícím VŠKP se ve formátu EVSKP-MS zapisuje jménem instituce jako volný text, nebo pomocí vnořeného prvku ccz:universityOrInstitution podle standardu CorpCZ.

V textové formě může být jméno instituce zpřístupňující VŠKP mapováno na opakovatelný prvek formátu DC Set dc:publisher. Ve formátu PTS není odpovídající prvek a dochází tak ke ztrátě informace.

Datum doručení metadatového záznamu do repozitáře

Datum doručení metadatového záznamu do externího repozitáře z lokálního registru nemá jednoznačnou obdobu v dalších zmiňovaných metadatových formátech.

Zpřístupnění souborů VŠKP

Datum zpřístupnění veřejnosti nebo časový interval, ve kterém mohou být soubory eVŠKP zpřístupněny, je ve formátu PTS mapován na prvek pts:available vnořený v prvku pts:presentation.file (viz výše *Popis konkrétního souboru VŠKP*).

Ve formátu DCMI Terms je prvek dcterms:available identický, při mapování do formátu DC Set by při použití dc:date došlo k významné ztrátě sémantického významu a proto mapování prvku zde není doporučováno (pokud je prvek požadován, např. při harvestování metadat do centrálního indexu discovery služby PRIMO, je vhodné použít samostatný set pro harvestování s konkrétním nastavením dle specifických požadavků harvestující služby).

Datum změny záznamu VŠKP

Datum a popř. čas změny metadatového záznamu VŠKP v proprietárním prvku evskp:modified standardu EVSKP-MS je vhodné mapovat 1:1 na prvek formátu DCMI Terms dcterms:modified.

Při mapování do formátů PTS (prvek pts:getfile.modified) a DC Set (dc:date) by došlo k významnějšímu posunu sémantického významu a proto mapování není doporučováno, pokud to není výslovně požadováno provozovatelem harvestující služby.

4.3 Implementace OAI-PMH serveru na VŠE v Praze

Autor disertační práce implementoval v Databázi kvalifikačních prací VŠE OAI-PMH server, který je vzhledem ke standardizaci a rozšíření protokolu použitelný pro import dat do Národního úložiště šedé literatury, národního registru VŠKP Theses.cz, mezinárodního registru DART-Europe E-theses Portal a centrálního indexu discovery služby PRIMO. Níže uvedený referenční popis implementace OAI-PMH serveru na VŠE v Praze autor prezentoval na 3. ročníku semináře Systémy pro zpřístupňování VŠKP: zkušenosti, možnosti, nabídky, potřeby (66).

Pro implementaci OAI-PMH na Vysoké škole ekonomické v Praze bylo zvoleno prostředí PHP a databáze MySQL, která obsahuje metadata o eVŠKP obhajovaných na VŠE v Praze.

Požadavky a související odpovědi OAI-PMH je možné rozdělit na statické, kde odpovědí je převážně neměnný soubor XML, a dotazy dynamické, kde obsahem odpovědí jsou metadata generovaná z databáze. Základní URL adresa (tzv. Base URL) sloužící pro dotazování repozitáře VŠE je http://www.vse.cz/oai, jednotlivé parametry a typy dotazů jsou podle protokolu OAI-PMH uváděny v parametrech této URL adresy. Požadovaný příkaz pro OAI-PMH server je konkrétně specifikován v parametru verb URL adresy.

Ukázku implementace metadat eVŠKP v OAI-PMH serveru VŠE v Praze obsahuje Příloha V *OAI-PMH export metadat ve formátu Dublin Core* a Příloha VI *OAI-PMH export metadat ve formátu EVSKP-MS*.

4.3.1 Odpovědi statické

První kategorii dotazů tvoří identifikace serveru, dotaz na seznam podporovaných formátů metadat a dotaz na seznam metadatových sad záznamů.

Identify

Příklad: http://www.vse.cz/oai?verb=Identify

Odpověď v XML formátu obsahuje jméno repozitáře, e-mail administrátora, indikaci podpory smazaných záznamů v repozitáři aj. údaje o serveru.

List Metadata Formats

Příklad: http://www.vse.cz/oai?verb=ListMetadataFormats

XML záznam obsahuje seznam podporovaných metadatových formátů – prefix pro označení formátu, použité schéma XML a jmenný prostor formátu. Každý repozitář musí povinně poskytovat metadata v metadatovém formátu nekvalifikovaný Dublin Core (DC Set, metadatový prefix oai_dc), v případě exportu záznamů eVŠKP se použije metadatový formát EVSKP-MS (metadatový prefix oai_evskpms).

List Sets

Příklad: http://www.vse.cz/oai?verb=ListSets

Každý repozitář může interně členit záznamy do metadatových sad, které jsou označeny číslem nebo textem. Při požadavcích na metadata je možné určit, o jakou sadu máme konkrétně zájem. Na VŠE v Praze se pro eVŠKP používají následující sady:

- theses – vysokoškolské kvalifikační práce bez omezení typu
- bachelors_thesis - Bakalářské práce
- masters_thesis - Diplomové práce
- dissertations - Disertační práce

Další sady obsahují metadata článků z časopisů publikovaných na webu VŠE v Praze.

Rozdělení do sady souhrnné a do sad dílčích podle jednotlivých typů eVŠKP usnadňuje harvestování v případě, kdy je požadován export pouze specifických dat a harvestující repozitář nedokáže data jednoduše filtrovat. Příkladem může být např. registr DART-Europe E-theses Portal, harvestující pouze disertační práce ze sady dissertations, nebo centrální index discovery služby PRIMO indexující metadata eVŠKP a časopiseckých článků VŠE v Praze.

4.3.2 Odpovědi dynamické

Identifikátory dostupných záznamů poskytuje dotaz List Identifiers, konkrétní metadata potom dotazy List Records a Get Record.

List Identifiers

Příklad: http://www.vse.cz/oai
?verb=ListIdentifiers&from=2015-01-15&until=2015-01-20&metadataPrefix=oai_evskpms

URL adresa obsahuje kromě označení příkazu v parametru verb povinně metadatový formát, v jakém budeme chtít záznamy získat (v našem příkladu prefix oai_evskpms pro formát EVSKP-MS), nepovinně metadatovou sadu a časový rozsah, kdy mělo dojít k modifikaci metadat (ve formátu EVSKP-MS uloženo v prvku evskp:modified). Každý záznam v repozitáři je OAI-PMH serverem označen identifikátorem záznamu specifickým pro daný

server (nelze zaměňovat s trvalým identifikátorem EVSKP-MS uváděným samostatně v metadatech).

List Records

Příklad: http://www.vse.cz/oai
?verb=ListRecords&from=2015-01-15&until=2015-01-20&metadataPrefix=oai_evskpms

Na rozdíl od požadavku na identifikátory List Identifiers v této variantě dotazu získáváme přímo metadatové záznamy v požadovaném standardu (v příkladu EVSKP-MS). Metadatový záznam je uveden jako vnořený prvek v rámci prvku metadata, který se opakuje pro každý záznam splňující omezující podmínky v URL.

Get Record

Příklad: http://www.vse.cz/oai/
?verb=GetRecord&metadataPrefix=oai_evskpms&identifier=oai:vse.cz:vskp/4840

Pokud máme seznam identifikátorů, zaslaný jako XML v požadavku List Identifiers, můžeme záznamy stahovat po jednom – v URL adrese specifikujeme požadovaný prefix formátu metadat a identifikátor. Metadata záznamu jsou vložena opět v rámci prvku metadata, tentokráte se již na rozdíl od List Records tento prvek neopakuje.

4.3.3 Realizace exportu na VŠE

Základní naprogramování OAI-PMH serveru na VŠE v Praze trvalo přibližně 40 člověkohodin včetně studia dokumentace, převedení EVSKP-MS záznamů z verze formátu 0.1 na verzi 1.1 a odladění validity XML. Pro aplikaci bylo využito stávající Databáze kvalifikačních prací VŠE implementované za využití PHP a MySQL.

Základní adresa OAI-PMH serveru http://www.vse.cz/oai je obsluhována jedním PHP souborem, který dle typu požadavku využívá šest dalších souborů.

Stahování plných textů eVŠKP externí službou umožňuje prvek evskp:transfer ze standardu EVSKP-MS, který obsahuje URL adresy jednotlivých souborů eVŠKP.

Funkčnost byla testována pomocí online aplikace Repository Explorer http://re.cs.uct.ac.za. Získané XML záznamy, konkrétně obsah prvků evskp:metadata, byly kontrolovány

validátorem na serveru http://validator.nu oproti Relax NG schematu zpracovanému pro poslední verzi EVSKP-MS standardu.

Vzhledem k požadavku provozovatele centrálního indexu discovery služby PRIMO, aby VŠE v Praze uváděla v metadatovém formátu DC Set (metadatový prefix oai_dc) prvek dcterms:available (datum, od kdy je plný text dostupný v případě embarga), bylo nutné přizpůsobit OAI-PMH server univerzity. Prvek dcterms:available není součástí metadatového standardu DC Set, ale součástí DCMI Terms. Použití prvku tak způsobovalo chybnou validaci metadat u většiny externích služeb nabírajících metadata z repozitáře VŠE v Praze. Prvek dcterms:available je proto uváděn pouze v případě přístupu ze specifických IP adres OAI-PMH klienta discovery služby PRIMO. V případě dalších OAI-PMH klientů, z odlišných IP adres, není prvek dcterms:available v poskytovaných metadatech uveden a metadata jsou tak validní podle DC Set.

Posledním klientem, pro kterého byl OAI-PMH server zásadněji upravován, je evropský portál DART-Europe E-theses Portal - DEEP (viz oddíl 2.3.1). Podmínkou zařazení do indexu DEEP jsou:

a) zpřístupnění plných textů v režimu Open Access,
b) zpřístupnění vědeckých, odborných kvalifikačních závěrečných prací,
c) podepsání Dohody o partnerství knihoven s cílem vytvořit jednotný evropský portál.

Po zveřejnění plných textů z Databáze kvalifikačních prací VŠE v režimu Open Access, oslovil autor zástupce DEEP[24] se žádostí o začlenění repozitáře VŠE v Praze do indexu portálu DEEP. Podmínka nabídky vědeckých, odborných kvalifikačních závěrečných prací byla splněna vytvořením specifických setů podle jednotlivých typů eVŠKP – portál DEEP konkrétně harvestuje set dissertations obsahujícího pouze metadata disertačních prací.

Metadata eVŠKP ve formátu EVSKP-MS jsou na straně OAI-PMH serveru mapována na prvky DC Set, metadatový prefix oai_dc (viz doporučené mapování výše). Na VŠE v Praze se jedná o prvky dc:title (opakující se prvek, např. český název práce a anglický překlad),

[24] Jednání probíhala konkrétně se správcem portálu Martinem Moyle z University College London, který na pozvání autora již v ČR na téma DEEP přednášel v rámci 4. ročníku semináře Systémy pro zpřístupňování VŠKP: zkušenosti, možnosti, nabídky, potřeby: 4. ročník semináře konaného 21. 10. 2009 na VUT v Brně (40).

dc:creator, dc:subject, dc:description, dc:publisher, dc:contributor (opakovatelný prvek – vedoucí práce, oponenti), dc:date (datum obhajoby), dc:type (český a anglický termín označující volným textem typ práce, doplňkově identifikátor podle slovníku *OpenAIRE Guidelines* (63)), dc:format, dc:identifier (URL odkaz na HTML popis záznamu v Databázi kvalifikačních prací VŠE), dc:language a dc:rights (opakovatelný prvek, práva česky a anglicky).

Po podpisu dokumentu *Dohoda o partnerství knihoven s cílem vytvořit jednotný evropský portál* (použitý překlad autora viz Příloha II) zástupci VŠE v Praze a DART-Europe bylo v březnu 2014 započato pravidelné nabírání metadat z OAI-PMH serveru VŠE v Praze. Jan Mach se stal zástupcem VŠE v Praze v radě DART-Europe.

4.4 Závěr kapitoly

Na základě analýzy jednotlivých metadatových formátů autor disertační práce vypracoval doporučené mapování prvků EVSKP-MS do formátů PTS, ETD-MS, jednoduchý nekvalifikovaný Dublin Core (DC Set) a DCMI Metadata Terms (DCMI Terms).

Mapování je nutné v případě využití poskytování metadat protokolem OAI-PMH, při komunikaci metadat mezi institucemi, které nepodporují stejnou metadatovou sadu eVŠKP (typicky zahraniční nebo obecně zaměřené vyhledávače).

Analýza ukázala, že mapování do standardů Dublin Core a ETD-MS je možné bez obtížnějších transformací. Zásadnější ztráta informací je pouze při mapování prvků DCMI Terms (prefix dcterms) na prvky DC Set (prefix dc), kdy dochází ke ztrátě sémantické informace vzhledem k obecné definici prvků jednoduchého Dublin Core (např. dcterms:created nebo dcterms:dateSubmitted → dc:date).

V případě mapování na formát PTS doporučovaný Theses.cz je potřeba komplikovanějších transformací, kdy jeden prvek je možné mapovat na více různých prvků (např. dc:creator → dc:creator nebo pts:creator; evskp:fileProperties → pts:presentation.file + prvky vnořené), nebo je potřeba rozhodnutí o vhodném zdrojovém prvku pro mapování (např. výběr vhodného dostupného metadatového prvku z EVSKP-MS pro mapování na povinný prvek dcterms:dateSubmitted v PTS).

Vzhledem k tomu, že často dochází k zobecnění významu prvku použitého v EVSKP-MS (např. evskp:modified → dc:date), nelze automaticky toto doporučení aplikovat i v obráceném směru (např. dc:date → evskp:modified).

Na konci ledna 2015 bylo z Databáze kvalifikačních prací VŠE přes protokol OAI-PMH dostupných přes 37 000 záznamů eVŠKP ve formátu EVSKP-MS. Záznamy jsou sklízeny OAI-PMH klienty např. Theses.cz, NUŠL, PRIMO nebo DART-Europe E-theses Portal. V lednu 2015 portál DEEP obsahoval přes 560 záznamů disertačních prací VŠE v Praze, včetně odkazů na volně dostupné plné texty. VŠE v Praze je prozatím jedinou institucí z ČR pravidelně poskytující metadata do portálu DEEP. Jan Mach je členem rady DART-Europe.

5 Metriky pro měření užití eVŠKP v online prostředí

Cílem kapitoly je poskytnout přehled a kriticky zhodnotit, pro potřeby repozitářů eVŠKP, vhodnost metrik používaných při klasickém publikování a metrik používaných v prostředí WWW.

Při vědomí významu tradičních používaných bibliometrik pro hodnocení vědy není cílem stávající citační metriky a jejich význam zavrhovat, ale kriticky zhodnotit nové metriky a postupy, jejichž případné využití může pomoci k lepší evaluaci otevřených repozitářů eVŠKP a jejich obsahu a tím potažmo poskytnout podpůrnou argumentaci pro jejich další rozvoj a využití. V textu mj. ověříme hypotézu, zda alternativní metriky založené na analýze ohlasu dokumentů v sociálních sítích mají významný přínos pro hodnocení eVŠKP v českých repozitářích.

Na základě analýzy stávajících projektů autor v závěru kapitoly připraví doporučení pro agregaci a výpočet užití eVŠKP, distribuovaných v rámci repozitářů v ČR.

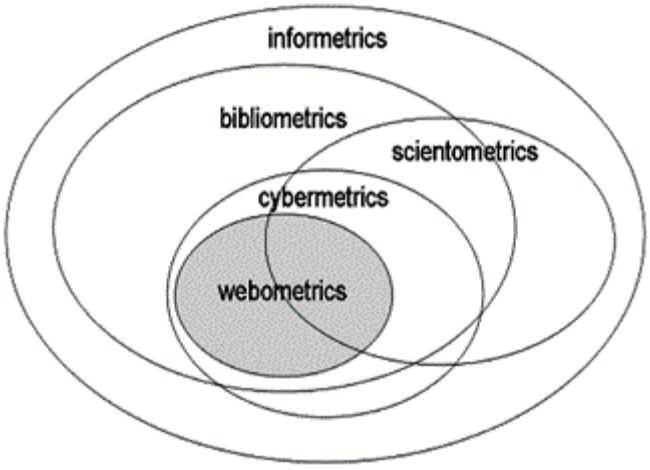
Tato kapitola je rozšířením práce *Metriky pro Open Access repozitáře* (67), vypracované v srpnu 2012 pro atest z předmětu Informační věda. Příspěvek *Statistiky využití článků v online repozitářích,* vycházející z této kapitoly, byl přijat na konferenci INFORUM 2015, která se bude konat v květnu 2015 v Praze.

5.1 Úvod do metrik pro repozitáře

Vzhledem k prosazování Open Access publikování v otevřených repozitářích (např. ArXiv, NUŠL, v institucionálních repozitářích aj.) a Open Access časopisech (např. časopisy PLOS) vyvstává potřeba vhodných metrik pro kvantifikaci vědeckých informací, dostupných online v prostředí Internetu. Historicky byly vědecké práce spojeny s konkrétním časopisem, nyní jsou však často dostupné samostatně online a mohou být přečteny a využívány nezávisle na dostupnosti časopisu a jeho reputaci. Vědci již nečtou celé časopisy jako celek, ale elektronicky vyhledávají odpovídající články napříč více zdroji, korelace mezi významem časopisu a citovaností článku klesá (68). Podobně vysokoškolské kvalifikační práce již většinou nejsou uzavřeny do knihoven a repozitářů univerzit, ale dochází k jejich šíření v elektronické podobě do národních a mezinárodních portálů.

5.1.1 Přehled souvisejících oborů

Přehled oborů zaměřených na komunikaci a měření vědeckých poznatků a vzájemný vztah oborů znázorňuje Obrázek 7.



Obrázek 7 Obory zaměřené na komunikaci vědeckých poznatků (115)

Infometrie je disciplínou informační vědy a využívá aplikování matematických metod na obsah informační vědy.

Bibliometrie studuje formální aspekty textů, dokumentů, knih a informací, poskytuje metody pro jejich kvantitativní analýzu především formou citační a obsahové analýzy. Citační indexy, jako např. Web of Knowledge nebo Scopus, nám umožňují studovat vztahy mezi dokumenty dané citačními ohlasy – kdo, koho citoval, jaké články. Z těchto dat lze následně posuzovat popularitu a vliv autorů, článků a publikací. Mezi další metody bibliometrie patří např. tvorba thesaurů, frekvenční analýza termínů, zkoumání gramatické a syntaktické struktury textů, měření užití článků podle čtenosti, kvantifikace podle roku publikování, typu dokumentu (recenzovaný článek, monografie aj.), instituce, oboru zájmu aj.

Scientometrie je „věda o vědě“ – vědecká nauka zabývající se vědou samotnou, produkcí vědeckých informací a souvisejícími makro ukazateli.

Za zakladatele scientometrie je považován Eugene Garfield zabývající se indexováním vědeckých informací a citací a jejich využitím pro hodnocení vědy. Za více jak 50 let existence scientometrie vzniklo množství dobře zpracovaných vědeckých metod, nástrojů a podpůrných databází (v zahraničí např. databáze ISI – Thomson Scientific, Scopus, pro

lékařské obory PubMed, v České republice důležité ASEP a RIV), zformovala se rozsáhlá vědecká komunita, množství vědeckých institucí, pořádají se pravidelné konference, výstupy jsou publikovány mj. v prestižním časopisu Scientometrics.

Konkrétněji předmětem scientometrie je podle (69) zkoumání vývoje vědy samotné, vztahy mezi jednotlivými vědeckými disciplínami, kvantitativní hodnocení vědeckých výstupů a komunikace ve vědě, evaluace vědců, vědeckých institucí a jejich spolupráce.

Scientometrie se především zabývá

- kvantitativním měřením projektů, vědců, zdrojů financování, publikací, patentování, citování (podle instituce, země, jazyka, spoluautorství, tematické oblasti aj.),
- výzkumem jednotlivců, institucí, vědeckých komunit,
- identifikací vztahů mezi vědeckými disciplínami a vztahů mezi vědeckými komunitami,
- vývojem vědeckých disciplín, identifikací nových trendů.

Předmětem zájmu ve scientometrii jsou tedy vědci, instituce, vědecké programy, a dále s nimi související vědecké výstupy – realizované projekty, registrované objevy, patenty, publikace a jejich citace.

Webometrie je mladá vědecká disciplína aplikující infometrické a bibliometrické metody na prostředí celého Internetu, kde studuje kvantitativní aspekty vytváření a využití informačních zdrojů. Metriky používané ve webometrii jsou velmi důležité pro Open Access zdroje, které jsou nad rámec klasické tištěné produkce šířeny právě v prostředí WWW, a jejichž plný vědecký význam nemůže být kvantifikován tradičními metodami scientometrie. Zatímco termín webometrie, webometriky je používán spíše pro materiály v prostředí WWW, o trochu obecnější věda kybermetrie se zabývá obsahem a komunikací v kyberprostoru.

5.2 Metriky založené na počtu citací

5.2.1 Journal Impact Factor

Dříve, než se podíváme na metriky související s konkrétní publikací a autorem, je potřebné specifikovat indikátor Journal Impact Factor (zkráceně JIF). Tento indikátor citovanosti

časopisu poprvé zmínil Eugene Garfield v roce 1955. Garfield sám později stál při vzniku Science Citation Indexu (SCI) využívajícího právě indikátoru JIF.

Měření úspěšnosti časopisu jen na základě počtu citací může být nevyhovující, neboť prosté srovnání podle počtu citací za dané období může znevýhodnit časopisy s menším množstvím článků, ale s velkým významem (70). Navržený indikátor (71) se vypočítá pro daný rok *y* podle Rovnice 1:

$$JIF_y = \frac{počet\ citací\ (k\ danému\ roku\ y)\ prací\ publikovaných\ v\ uplynulých\ 2\ letech}{počet\ (citovatelných)\ článků\ publikovaných\ časopisem\ v\ uplynulých\ 2\ letech}$$

Rovnice 1 Journal Impact Factor pro rok y

Vzhledem k vývoji citovanosti časopisů se JIF počítá pro sledované časopisy pro každý konkrétní rok, je pak možné sledovat trendy ve vývoji JIF daného periodika v čase.

Na JIF má vliv obor výzkumu, citační hustota a poločas citovanosti (70). Citační hustotou je chápán průměrný počet citací, které získá článek v daném oboru, poločas citovanosti k danému roku udává, po kolika letech se objeví 50 % všech citací na články daného časopisu v citačních rejstřících (72).

Pro výpočet se využívá počtu citací za období dvou let. V případě kratšího období by indikátor vypovídal více o rychle se měnících vědních oborech (nízký poločas citovanosti). Jak by se změnilo pořadí časopisů podle JIF v případě změny tohoto období? Garfield v dřívějších výzkumech empiricky prokázal, že pořadí časopisů ze stejného oboru by se zásadně neměnilo, ale v případě časopisů z různých oborů může dojít k výrazným změnám. Např. pro obor fyziologie uvádí: „Když studujeme časopisy napříč obory, ohodnocení pro časopisy fyziologie se výrazně zvyšuje s růstem počtu let, ale ohodnocení časopisů v rámci dané kategorie se významně nemění.“ (71 str. 90).

Jedním z argumentů proti použití JIF je riziko negativních a nesprávných citací – odkazování na chyby v článku, doporučení editorů citovat články ze stejného časopisu (z důvodu zvýšení JIF časopisu) apod. Tyto případy však neovlivňují JIF zásadně, výrazněji se mohou projevit problémy se zkratkami afiliací, které jsou však podle Garfielda každý rok odstraněny ještě před vydáním *Journal Citation Report* (70). V praxi se však potvrzuje, že ve Web Of Science se tyto problémy i nadále objevují. Řešením je např. systematické dohledávání špatných citací knihovníky a jejich oprava ve Web Of Science, jak se děje např. na VŠE v Praze a na UK.

Jako další možný problém ve výpočtu JIF kritici uvádí započítání do vzorce JIF citací všech článků vydaných časopisem (část nad lomítkem), včetně komentářů, korespondence, zpravodajství, editoriálů, rozhovorů aj. textů (70, 73), které se však nezapočítávají mezi vědecké články (část pod čarou), navíc kategorizace článku je ovlivněna volbou lidí provádějících klasifikaci. V těchto případech by tak byl uměle zvýšen indikátor JIF. Tyto články jsou však podle Garfielda většinou citovány ve stejném roce a proto tento problém se týká jen malého množství hlavních časopisů (citovanost v daném roce se měří indexem bezprostředního vlivu nebo odezvy, někdy označován také jako Garfieldův index (72)).

JIF je ovlivněno velikostí vědního oboru, kterým se daný časopis zabývá. I zde lze uplatnit Parretovo pravidlo, že 20 % článků je odpovědných za 80 % citací. Pokud je vědní obor široký, publikuje zde mnoho autorů a jen málo z nich získává citace (70). Na JIF má větší význam citační hustota a stáří citované literatury, jak bylo zmíněno výše, nelze však zanedbat meziroční variabilitu JIF, viz (73) níže.

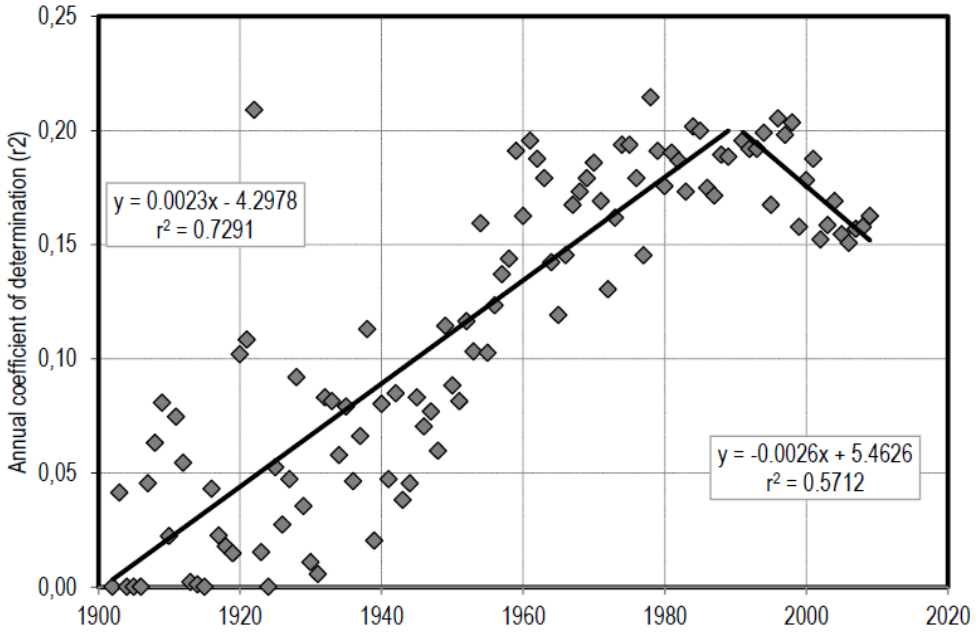
Autoři využívají JIF pro výběr časopisu, ve kterém by publikovali, příp. při rozhodování zda publikovat v konkrétním časopisu. Platí předpoklad, že časopisy s vysokým JIF jsou nejvíce prestižní, proto i JIF a potažmo počet získaných citací je jedním ze základních kritérií hodnocení vědy v ČR. Nově publikované články nemusely však citace za krátkou dobu, požadovanou pro hodnocení, získat, JIF je proto chápán jako očekávaný počet citací. Články, které jsou kandidáty na vysoký počet citací, uvádí Thomson Scientific jako *Hot Papers* v publikaci *Science Watch*, potvrzení úspěšnosti je však možné až po dvou letech od publikování po spočítání skutečných citací článku.

Při srovnávání časopisů podle JIF je tedy důležitý obor, ve kterém časopis publikuje. Někteří editoři by rádi porovnávali jen na základě nejcitovanějších článků, jiní doporučují řazení podle geografického nebo jazykového hlediska. Určitým řešením může být databáze *Journal Performance Indicators* (JPI). Oproti databázi *Journal Citation Report* „databáze JPI propojuje každý zdroj na jeho vlastní citace. Proto výpočty impaktu jsou přesnější. Jsou zahrnuty pouze citace podstatných položek, které jsou ve jmenovateli.“ (70).

Nejen Garfield v pracích (70, 74) upozorňuje na nutnost neustálé evaluace impakt faktoru, objevuje se exponenciální nárůst zájmu o tuto problematiku. Impakt faktoru se věnovalo v letech 1999 – 2007 přes 1100 záznamů nalezených ve Web of Science (60 % publikováno v posledních šesti letech). Garfield navrhuje zpracovat souhrnný kritický přehled literatury

publikovaný na téma JIF, což považuje za vhodný námět na jednu či více disertačních prací. Důvodem k sledování problematiky JIF je podle Garfielda také možnost nahrazení časopisů Open Access přístupem a repozitáři, upozorňuje na význam webometrik (kterým se budeme věnovat v souvislosti s eVŠKP v druhé části kapitoly): „Pokud k tomu dojde, ‚evaluace časopisů' sama o sobě může zaniknout, ale dýchejte zhluboka. Přicházejí výzkumné fronty, klastry spolu-citování, sémantické kategorie, hodnocení stránek a ostatní formy klasifikace, čímž ve své podstatě tradiční časopisy jsou." (74)

Korelace mezi množstvím citací článku a impakt faktorem časopisu (viz Obrázek 8), ve kterém byl publikován, roste v průběhu 20. století a klesá po roce 1990 s nástupem digitálního věku, jak dokazuje studie (68).



Obrázek 8 Koeficient determinace r2 mezi IF časopisů z fyziky a počtem citací za 2 roky článků publikovaných v nich v letech 1902 až 2002 (68)

Studie dále prokazuje, že „podíl 10 % nejcitovanějších prací publikovaných v 10 % nejcitovanějších časopisů klesá od roku 1990, z 5,25 % na 4,50 %. V souladu s tím, podíl 10 % nejcitovanějších prací nezveřejněných v 10 % časopisů s nejvyšším impact faktorem od roku 1990 roste, z 52 % na cca 56 %. Tento vývoj je ještě zřetelnější, když je stejné srovnání zpracováno pro horních 5 % prací a horním 5 % časopisů." (68 str. 10). Vzhledem k závěrům, že citace se začínají více rozprostírat mezi jednotlivé časopisy, studie predikuje,

že digitální věk a metody šíření a zpřístupňování vědeckých zdrojů mohou potlačit důležitost impakt faktoru jakožto významného kritéria důležitosti vědeckých publikací.

5.2.2 Citační ohlas

Hodnocení využití a dopadu konkrétních článků a autorů se zabývá citační analýza, konkrétněji indikátor citační ohlas nazývaný také citační index (termín citační index je možné chápat také jako databázi citačních ohlasů). Citační ohlas konkrétního vědeckého článku lze spočítat jako celkový počet citací daného článku[25]. Pokud bychom sečetli počet citací publikací konkrétního autora, získáme citační ohlas autora (uvádí se průměrný nebo celkový počet citací autora), podobně lze spočítat citační ohlas konkrétního časopisu (průměrný počet citací pro články v časopisu viz oddíl 5.2.1 *Journal Impact Factor* výše).

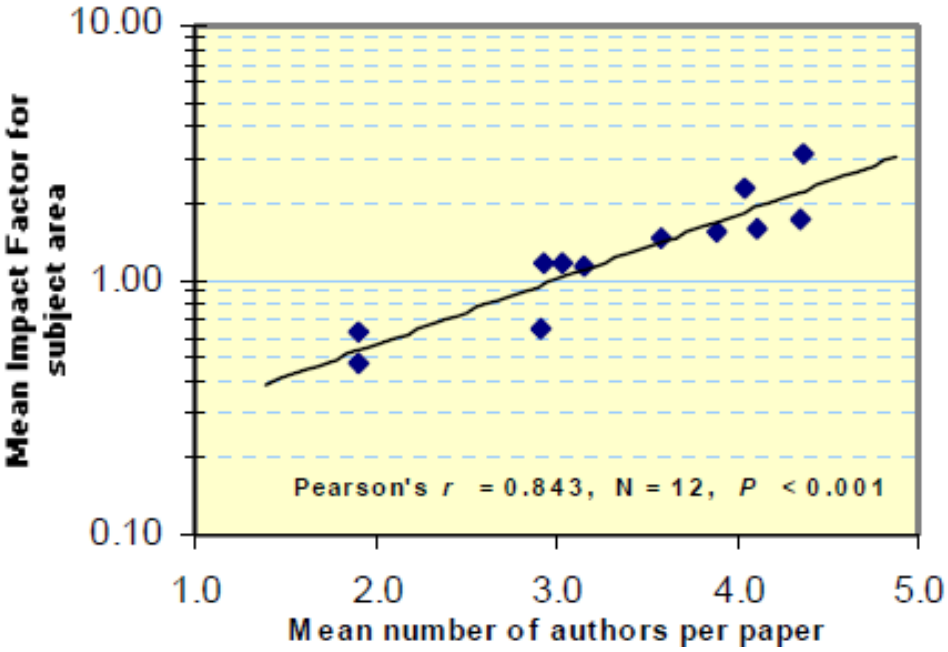
Typicky se pro výpočet počtu citací využívá citačních databází. První databáze citací v akademických časopisech byly spravovány Institutem vědeckých informací (Institute for Scientific Information, databáze Web of Knowledge), vznikaly další oborově zaměřené (Scopus společnosti Elsevier, NASA ADS aj.). Až na konci 20. století, s nástupem automatizace, došlo k nástupu automatické tvorby těchto databází díky extrakci textu – první citační databáze CiteSeer v roce 1997, později Google Scholar, Microsoft Academic Search aj. Jednotlivé databáze se liší zaměřením, mírou pokrytí a přesností (více záznamů, především automatizovaně získaných, determinuje nižší přesnost). Např. Web of Knowledge má dobré retrospektivní pokrytí časopiseckých publikací, ale špatné pro velmi významné konference a jejich publikace; Scopus má pokrytí konferencí lepší, ale podobně jako Google Scholar obsahuje záznamy spíše až od 90. let minulého století. Google Scholar obsahuje výrazně více záznamů než předešlé databáze, ale jedná se často o záznamy citací v šedé literatuře, které výrazněji nepřispívají ke zvýšení vědecké úrovně autora.

Platí předpoklad, že úspěšné články a úspěšní autoři mají větší počet citací. Citační ohlas je možné použít jako jedno z mnoha kritérií hodnocení vědecké práce, není vhodné ho ale použít jako jediné či hlavní kritérium hodnocení vědecké úrovně, kvality či významu.

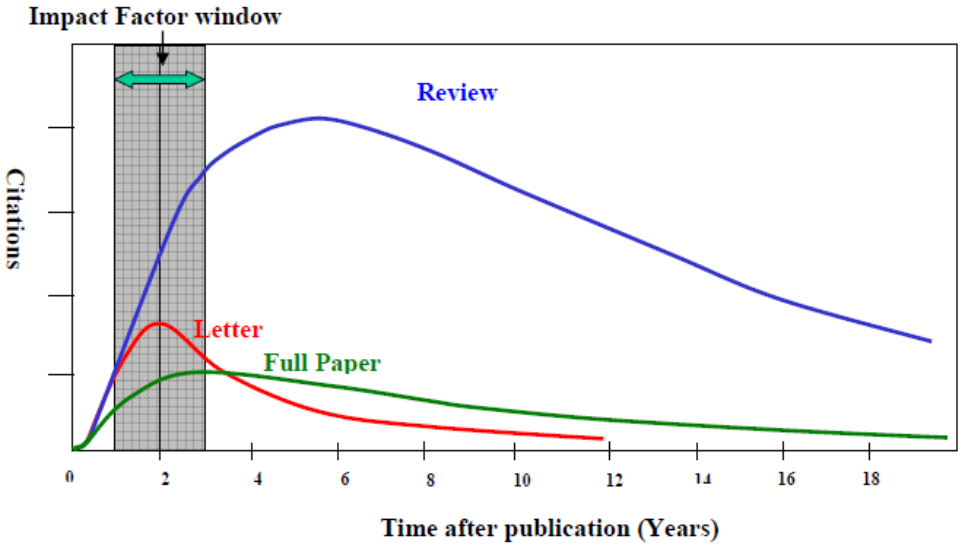
[25] Citací je chápán proces reference (použití) či citování autora, roku, titulu práce a zdroje publikování jakožto zdroje použitého při psaní dané odborné práce.

Varování před nevhodným užitím impakt faktoru a jeho problémy jsou uvedeny v mnoha pracích, důkladně se jim věnuje např. článek *Impakt faktory: využívání a zneužívání* autorů ze společnosti Elsevier (73). Z mnoha studií zkoumaných při přípravě této práce vyplývá, že hodnocení jen na základě množství citací (impakt faktoru) nevypovídá přesně o kvalitách autora, neboť impakt faktor nezohledňuje mnohé sociologické a statistické faktory - typ článku (krátká oznámení, přehledové články, odborné články o novinkách …), oborové zaměření časopisu, velikost časopisu (větší časopisy mají menší fluktuaci JIF v jednotlivých letech) a počet vydání ročně, počet spoluautorů (souvisí s vědním oborem), autocitace, negativní citace, autory s velkým množstvím článků s malým počtem citací, délku sledovaného období citování a poločas citovanosti apod. faktory. Impakt autora či instituce nelze měřit prostým vynásobením počtu článků v jednotlivých časopisech a jim odpovídajících JIF, neboť jen malé množství článků generuje většinu citací pro konkrétní časopis, JIF pracuje ale s průměrem (75).

Studie (73) dokládá variabilitu JIF na jednotlivých faktorech, např. mezi oborem výzkumu a JIF časopisů (nejlepší časopisy z jednoho oboru mohou mít JIF horší než málo kvalitní v jiném oboru (viz Obrázek 9). Autoři často citují sebe navzájem, což vede k odlišnosti běžného počtu spoluautorů v jednotlivých vědách. Časopisy s krátkými články a oznámeními mají vyšší počet okamžitých citací, ale oproti recenzovaným časopisům krátký poločas citovanosti (viz Obrázek 10).



Obrázek 9 Vztah mezi průměrným impakt faktorem v oboru a počtem autorů (73)



Obrázek 10 Vztah mezi počtem citací a dobou po publikování podle typu časopisu (73)

Z výše uvedených studií vyplývá, že je nepřípustné použít impakt faktor pro hodnocení všech časopisů (i s odlišením podle kategorie článků) ve všech oborech či u individuálních autorů, využití impakt faktoru má své limity, včetně využití „oficiálního“ JIF od ISI. A to nejen kvůli výše uvedené variabilitě, ale také kvůli trendům v jednotlivých oborech a změnách oborů v čase.

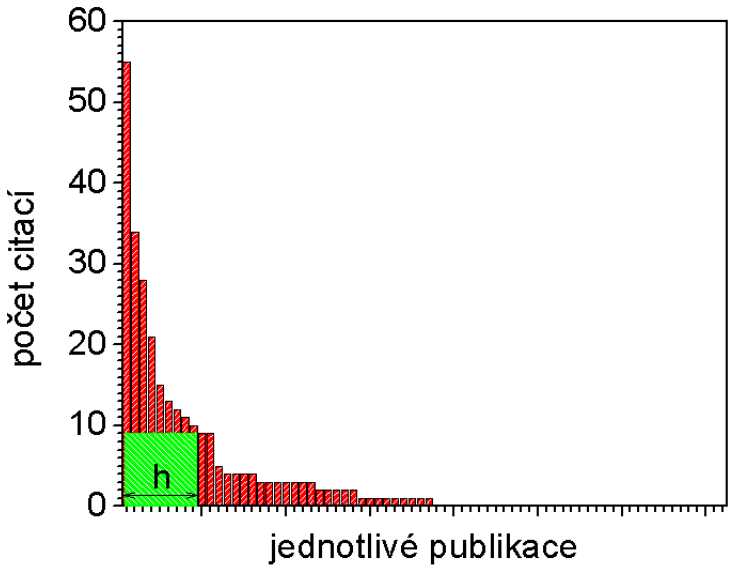
Pro účely hodnocení vědeckých pracovišť a týmů je proto vhodné kombinovat bibliometrické metriky odvozené z citační analýzy s dalšími faktory, jako jsou velikost rozpočtu, zdroje příjmů, recenzní řízení aj. Metriky na bázi citační analýzy se prozatím vyvíjejí, spíše než aby byly nahrazeny metrikami jinými. Problémy klasického impakt faktoru se snaží řešit metriky odvozené z počtu citačních ohlasů, nejvýznamnější je Hirchův index zmíněný níže. Výzkumy alternativních metrik existují (viz dále), ale prozatím nejsou tyto metriky při hodnocení vědy adekvátní, oficiálně užívanou, náhradou metrik na bázi citační analýzy.

„Měření citací, usnadněné bohatostí citačních databází ISI, může poskytnout velmi užitečný vhled do vědeckého výzkumu a jeho komunikace. Impakt faktor, jako jedna z citačních metrik, je užitečný při určování vlivu časopisů v rámci literatury daného oboru. Nicméně, není přímým měřením kvality a musí být používán se značnou opatrností.“ (73 str. 6)

5.2.3 Hirschův index a metriky odvozené

Hirschův index (zkráceně h-index, h-number) zohledňuje zároveň jak produktivitu, tak i význam publikovaných prací autora na základě distribuce počtu citací u nejvíce citovaných článků autora. Metriku navrhl Jorge E. Hirsch pro měření relativní kvality teoretických fyziků. Tato metrika je v současnosti velmi významná, používaná pro hodnocení autorů, je na ni např. kladen důraz při rozhodování o grantových projektech v ČR.

Vědec má index ve výši h, pokud h jeho N_p článků má nejméně h citací každý, a ostatních (N_p-h) článků nemá více než h citací každý (viz Obrázek 11). Původní definice Hirsche uváděla podmínku, že ostatní články mají mít méně než h citací, což v určitých situacích nemůže být splněno (76).



Obrázek 11 Setříděný histogram počtu citací publikací daného autora s grafickým vyznačením významu h-indexu (116)

Výpočet může být prováděn ručně nebo za využití citačních databází, které provádějí výpočet automaticky. Výše h-indexu se ale liší podle použité databáze a množství zaindexovaných záznamů. Např. Harzingův program Publish or Perish počítá h-index podle databáze Google Scholar, která obsahuje více citací, ale s menší přesností (s větší chybovostí) než citační indexy Web of Knowledge a Scopus.

Metrika h-index odstraňuje některé, i když ne všechny výše uvedené nedostatky citačního indexu. Eliminuje vliv velkého množství prací autora jen s malým počtem citací (a je proto jednoduchá na výpočet), potlačuje efekt (spolu)autorství jen na jedné práci s velkým počtem citací (dané např. typem publikace, viz kritika citačních ohlasů výše), které nezohledňuje autorovu vědeckou kvalifikaci.

Podobně jako u dalších bibliometrik, h-index může být nesprávně interpretován vlivem mnoha faktorů, jako je např. množství spoluautorů, odlišná distribuce počtu citací mezi publikovanými články autora (řeší g-index (77)), průměrný počet citací v jednotlivých oborech, roli spoluautora, délku publikační činnosti (s délkou vědeckého výzkumu autora a souvisejícím počtem publikací roste h-index), bez ohledu na význam aktuální práce (pro započtení vlivu lze kombinovat s Eigenfactorem, viz níže), důvod citace (poděkování, autocitace, automaticky generované dokumenty chybně indexované v Google Scholar aj.), typ publikace (např. v humanitních vědách se více publikují knihy – toto bere v potaz až např. bodové hodnocení vědy dle RIV v ČR).

Výše uvedené nedostatky se snaží řešit metriky odvozené z h-indexu, jako jsou např. normalizovaný h-index (odstranění některých mezioborových rozdílů), A-index a g-index (započítání vlivu množství citací při zachování vhodných vlastností h-indexu), m-index (podíl h-indexu a počtu let publikační činnosti), c-index (zahrnující spoluautorství, resp. min. „vzdálenost“ spoluautorů) či Succesive Hirsch-type-index měřící úspěšnost instituce (roven *i* tehdy, pokud alespoň *i* vědců má h-index ve výši *i*).

U jednotlivých indikátorů, které se liší dle různých aspektů publikační činnosti autora, je potřeba si přesně uvědomit, které vlivy zahrnují, v jaké míře, a které potlačují či přímo ignorují. „Právě vzhledem k mnoha omezením, které jednotlivé indikátory mají, nelze nahradit kvalitativní posuzování, stejně jako nelze brát výstupní hodnoty jednotlivých metrik jako absolutní bernou minci. K rozhodnutí o zapojení indikátoru jako hodnotícího prvku je nejprve nutné se důsledně seznámit s jeho vlastnostmi, definicemi a užitými metodami tak, aby alespoň co nejlépe odpovídal záměru hodnocení.“ (72 str. 11)

5.2.4 Eigenfactor

Indikátor Eigenfactor score navrhli v roce 2007 Jevin West a Carl Bergstrom pro měření celkové důležitosti vědeckých časopisů na základě získaných citací. Oproti JIF je zde započítána důležitost referujících časopisů, odkazující prestižní časopisy zvyšují Eigenfactor více než málo významné. Metrika tak odpovídá počítání Google PageRank, kde často odkazované weby (tedy chápejme důležité) odkazující na cílový hodnocený web přispívají výrazněji k hodnocení tohoto cílového webu a tím k lepší pozici hodnoceného webu ve výsledkové listině nalezených stránek na dotaz v Google.

Druhý indikátor zjistitelný prostřednictvím webu EigenFactor.org – „Article Influence“ – je určen k odhadnutí důležitosti časopisu nehledě na jeho velikost (počet publikovaných článků). Article Influence vyjadřuje míru průměrného vlivu každého článku v časopise za pět let po jeho publikování. (72).

Pokud indikátory Eigenfactor zkombinujeme s h-indexem, získáme indikátor důležitosti individuálního vědce.

5.2.5 Y-factor

Obdobou Eigenfactoru je Y-factor navržený Herbertem Van de Sompelem kombinující JIF s váženým PageRankem.

„Y-faktor zavádí do hodnocení časopisů pojem popularity a prestiže (…) Populární časopisy jsou ty, které jsou často citovány časopisy s nízkou prestiží. Tyto časopisy mají velmi vysoký impakt faktor a velmi nízký Vážený PageRank. Prestižní časopisy jsou ty, které nejsou tak často citovány, ale jejich citace pocházejí z vysoce prestižních časopisů. Tyto časopisy mají nízký impakt faktor a vysoký Vážený PageRank.“ (72 str. 16)

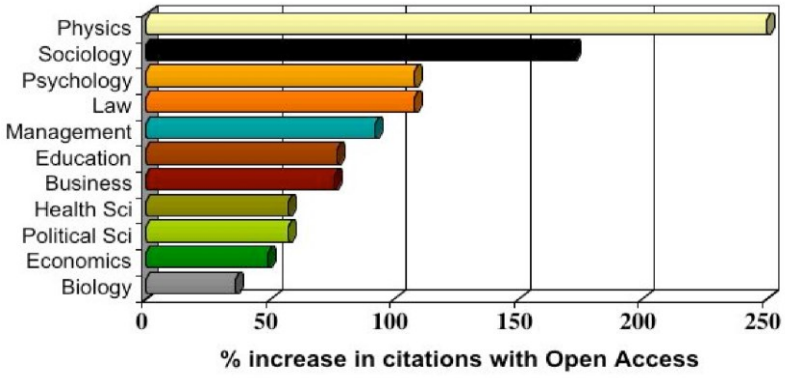
5.3 WWW a Open Access

Jak již bylo nastíněno výše, s nástupem World Wide Web se mění způsob vědecké komunikace, od klasického publikování v tištěných periodikách k šíření vědeckých informací v online prostředí Internetu, od publikování v časopisech kupovaných čtenáři k volnému přístupu placenému někým jiným než uživatelem – např. autorem, institucí apod.

Publikování vědeckých informací v režimu Open Access (viz oddíl 2.1.4) může mít různé podoby, Budapešťská iniciativa (12) definuje zlatou nebo zelenou cestu k Open Access, příp. může být i jejich kombinace. V případě zlaté cesty („golden roads to Open Access“) probíhá klasické podání příspěvku redakci časopisu, recenzní řízení a následné zveřejnění v časopisu s tím, že článek časopis poskytuje čtenářům zdarma ke stažení. Náklady na vydání v tomto případě neplatí čtenář časopisu (pokud si nekoupí tištěné vydání), ale hradí je sám autor nebo jeho mateřská instituce, příp. poskytovatel grantových prostředků. V případě zelené cesty („green roads to Open Access“) probíhá opět recenzní řízení, autor získává od vydavatele v rámci licence právo uložit preprint, postprint či tiskovou verzi své práce na svém webu či v institucionálním nebo oborovém repozitáři. V případě publikace preprintu nejsou zapracovány připomínky z recenzního řízení a při citování může dojít k převzetí chyby opravené v postprintu.

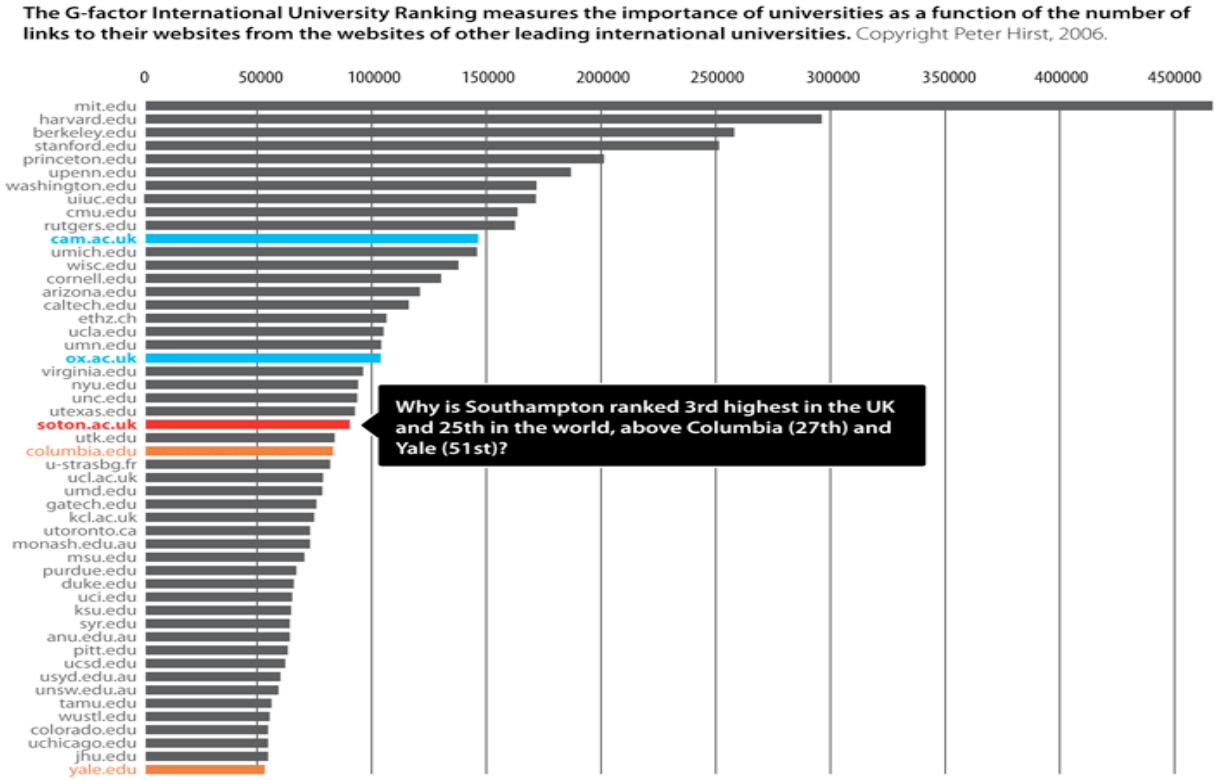
Nevýhodou uložení článku na vlastních WWW stránkách autora je větší pracnost, menší viditelnost/dohledatelnost článku, nízká interoperabilita a krátkodobá životnost takových stránek. Výhodou institucionálních repozitářů je možnost napojení na oborové a přehledové databáze (např. DRIVER, RePEC pro ekonomii, PhilSci Archive pro filozofii vědy, Scirus aj.)

Nárůst počtu citací vědecké práce při publikování v režimu Open Access dokládají ve svém článku Brody a Harnad (78), viz Obrázek 12.



Obrázek 12 Zvýšení počtu citací při publikování v režimu Open Access (78)

Ve studii (78) a dalších pracích Steva Harnada je dokládán vliv raného zavedení Open Access politiky (Open Access mandate) pověřující autory instituce autoarchivací plných textů v institucionálním repozitáři, konkrétně vliv na vysoké hodnocení G-factor International University Ranking dané množstvím získaných odkazů na vlastní webové stránky z webů významných zahraničních institucí (viz Obrázek 13).



Obrázek 13 Význam institucionálního repozitáře (78)

5.4 Webometrické indikátory

Vzhledem k uvedenému nástupu WWW a změně publikování se objevila potřeba nových metrik, které vezmou v potaz nové faktory úspěšnosti vědecké práce než jen pouhý počet citací v impaktovaných časopisech. Může se jednat o indikátory na bázi odkazů, kde odkaz na článek či eVŠKP v prostředí WWW odpovídá citaci v impaktovaném časopise (72), o indikátory známé z webometrie používané pro hodnocení WWW stránek, nebo nově o tzv. sociální metriky, které měří úspěšnost a odezvu na určitou zprávu v sociálních sítích.

Dokumenty na Internetu získávají širší počet čtenářů než ty publikované v tištěných periodikách. Díky sledování aktivit (nepublikovaných materiálů, draftů, prezentací, studijních materiálů, komentářů, bookmarkování apod.) vědeckých skupin, profesorů, postgraduálních studentů aj. osob na webu mohou vzniknout nové indikátory, postihující buď důležitost vědecké práce v odlišném ohledu než klasická citační analýza na bázi impakt faktoru, nebo kvality vědecké práce nepostihnutelné citační analýzou (práce kontroverzní nebo podněcující diskusi, horká témata, ekonomické, sociální, kulturní nebo oborové vztahy mezi autory apod.).

Z velkého množství webometrických indikátorů si v této kapitole představíme vybrané významné metriky a jejich indikátory, které nám mohou pomoci při hodnocení významu vědeckých publikací.

5.4.1 Počet odkazů

Indikátory založené na měření počtu odkazů v prostředí WWW na daný dokument patří mezi základní indikátory webometrie (např. Web Impact Factor Petera Ingwersena). Postihují především atraktivitu odkazovaných dokumentů než jejich význam. Lze počítat počet odkazů na konkrétní dokument (atraktivita dokumentu), součet/relativní podíl/průměrný počet odkazů na všechny online dokumenty autora (atraktivita autora) či na konkrétní doménu (atraktivita instituce, vědeckého týmu, repozitáře apod.).

Tyto metriky není doporučováno využívat, bez dalších úprav, pro nejednoznačnost jejich výpočtu a interpretaci. Chybovost může být dána vyhledávacím nástrojem (např. tzv. Google

dance – rozdílnost indexů na jednotlivých národních doménách Google), aproximací výsledků (např. bráno v potaz 100/200/500 prvních nalezených výsledků), irelevancí odkazujících webů (odkazující zdroje neprocházejí recenzním řízením, může se jednat o automaticky generované odkazy) aj.

V prostředí WWW často dochází k tomu, že daný dokument je dostupný z více zdrojů, např. záznam eVŠKP z repozitáře univerzity distribuovaný do repozitářů Theses.cz, NUŠL a následně OpenAIRE. Je proto potřeba důsledná deduplikace, kdy ne vždy je možné využít automatické analýzy duplicit, natož nelze spoléhat na dokonalost analýzy citací (viz např. Google Scholar). Tento problém by řešilo důsledné použití perzistentních identifikátorů typu DOI, urn:nbn, HANDLE, PURL aj. pro identifikaci zdrojů, příp. ještě lépe až sémantický web podchycující např. i vazby mezi dokumenty, autory apod. Sémantický web sice má již dobré teoretické základy, své standardy, ale aplikace těchto zásad v praxi je prozatím horší.

5.4.2 Viditelnost odkazů

Další identifikátory webometrie, které nám mohou pomoci, jsou založeny na měření provázání jednotlivých zdrojů odkazy – měří externí odkazy, diverzifikaci odkazů v rámci jednoho webu, váhu/význam jednotlivých odkazů (např. Google PageRank) aj. aspekty odkazování. Tyto identifikátory nacházejí praktické uplatnění především v kombinaci s dalšími ukazateli jako identifikátory důležitosti odkazujících stránek či při analýze provázanosti zdrojů.

5.4.3 Velikost, počet stran

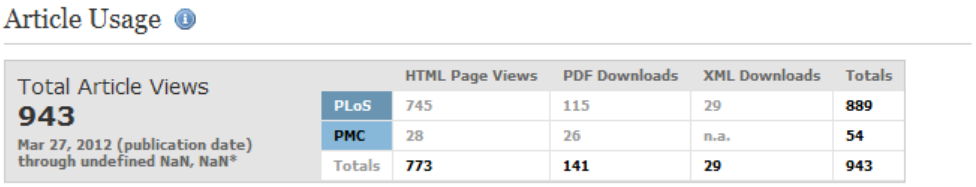
Další významná skupina webometrických indikátorů je založena na počítání velikosti webových sídel (počet WWW stran, subdomén, blogů …), plných textů (PDF dokumenty indikující možné akademické materiály, multimédia) či počet akademických článků v repozitářích.

Výše uvedené typy indikátorů již nacházejí své uplatnění v některých metodikách, především při hodnocení institucí (Ranking Web of World universities http://www.webometrics.info) či repozitářů (např. žebříčky OpenDOAR http://www.opendoar.org).

5.4.4 Počet stažení

Jedním z důležitých webometrických indikátorů je počet stažení („hit“), ať již stránky nebo dokumentu, např. eVŠKP. Počet stažení využívají v metrikách projekty jako např. COUNTER (viz oddíl 5.5.1), PIRUS (viz oddíl 5.6.2), LogEC aj. viz níže.

V praxi se odlišuje počet zobrazení vstupní HTML stránky s metadaty (např. jména autorů, název článku, abstrakt, klíčová slova), počet zobrazení plného textu (nejčastěji PDF), příp. i sémantické reprezentace dokumentu (např. XML, DC, RDF). Takto odlišené indikátory můžeme najít např. v rámci Open Access časopisu *PLOS ONE* (87) viz Obrázek 14.


Article Usage

Total Article Views
943
Mar 27, 2012 (publication date) through undefined NaN, NaN*

	HTML Page Views	PDF Downloads	XML Downloads	Totals
PLoS	745	115	29	889
PMC	28	26	n.a.	54
Totals	773	141	29	943

Obrázek 14 Metriky PLOS ONE založené na počtu zobrazení článku (87)

Rizikem při použití indikátoru tohoto typu je nebezpečí generování a následného započítání falešných návštěv, a to ať již těch nechtěných – prostřednictvím vyhledávacích robotů, nebo záměrně automaticky generovaných samotnými autory pro umělé navýšení návštěvnosti.

Pro správnou interpretaci metrik je proto nutné ze statistik odstranit přístupy robotů a opakované klikání jednoho uživatele (tzv. Doubleclick). Často se používá negativního seznamu IP adres vyhledávacích robotů, příp. statistický přístup. Jednotný způsob postupu však doposud neexistuje. (75)

Indikátor počet stažení je jednoduché získat ze statistik webového serveru, pouze je nutné při návrhu webu dodržet zásadu jednoznačnosti a neměnnosti identifikace cílového dokumentu (URL adresy) a vhodně eliminovat vliv falešných stažení.

Složitější je situace v distribuovaném prostředí, což jsou typicky eVŠKP dostupné z několika vyhledávacích služeb najednou. Cílem kapitoly je proto najít vhodné řešení výpočtu metrik užití eVŠKP aplikovatelné na zpřístupňování eVŠKP v ČR.

5.4.5 Návštěvnost, návštěvníci

Návštěvníkem je myšlena návštěva konkrétního uživatele po určitý časový úsek, tj. jeden návštěvník během své práce většinou stahuje ze stejného webu více dokumentů, příp. i stejný dokument opakovaně.

Nejednotnost měření počtu návštěvníků v jednotlivých repozitářích může být dána rozdílnou definicí časového úseku, což některé metodiky řeší jednotným, centralizovaným zpracováním institucionálním logů (např. Open Access Statistics (75)) nebo využitím centrálního logu (např. u benchmarkingu knihoven BIX (79) obsahuje cílová WWW stránka průhledný obrázek stahovaný z centrálního serveru, tj. logy jsou generovány a následně zpracovány jednou univerzitou). Pro knihovní prostředí je vhodné vycházet z definice virtuální návštěvy definované v *ISO 2789:2006 Information and documentation*, aby bylo možné porovnávat výsledky mezinárodně.

5.4.6 Sociální sítě, social bookmarking, citační manažery

Uživatelé mohou v internetových aplikacích pro tzv. social bookmarking vytvářet záznamy o webových stránkách podobně, jako se vytvářejí záložky oblíbených stránek v prohlížečích WWW. Tyto online záznamy je možné dále organizovat, anotovat a doplňovat vlastními klíčovými slovy.

Primárním účelem takovýchto aplikací bylo nejen umožnit online přístup k záložkám, ale také snazší vyhledávání WWW stránek za využití citační analýzy (počet záznamů jednotlivých URL adres, počet a frekvence jednotlivých klíčových slov přiřazených uživateli). Příkladem bookmarkovacích služeb jsou reddit.com (http://reddit.com), del.ico.us (http://www.delicious.com), digg (http://digg.com) aj.

Obdobný přínos jako bookmarkovací služby výše mohou pro nás mít sociální sítě na Internetu a jejich systém odkazování a hodnocení - blogy (odkazy, komentáře, hodnocení …), odkazování ve Wikipedii a aktivity v sociálních sítích Facebook (odkazy, sdílení, komentáře, Like) či Twitter (odkazy, followers, retweets).

Pro naši potřebu hodnocení eVŠKP a Open Access článků obecně mají pro nás o něco větší význam online citační manažery, které se více zaměřují na vědeckou komunikaci, využívány jsou spíše akademickou obcí. Kromě správy online zdrojů na WWW je možné vytvářet

i záznamy pro tištěné publikace a tyto záznamy dále kategorizovat. Příkladem citačních manažerů jsou RefWorks (http://www.refworks.com), EndNote (http://endnote.com), Mendeley (http://www.mendeley.com), CiteULike (http://www.citeulike.org) či Zotero (http://www.zotero.org).

Pokud získáme seznam záznamů konkrétní práce z uvedených aplikací (ať již identifikovaných podle URL, či přesněji podle DOI aj. identifikátorů), můžeme využít nástrojů citační analýzy podobně jako při měření citačního indexu pro tištěné publikace. Příkladem takového užití jsou metriky sociálních sítí v časopise *PLOS ONE* (viz Obrázek 18 na straně 131).

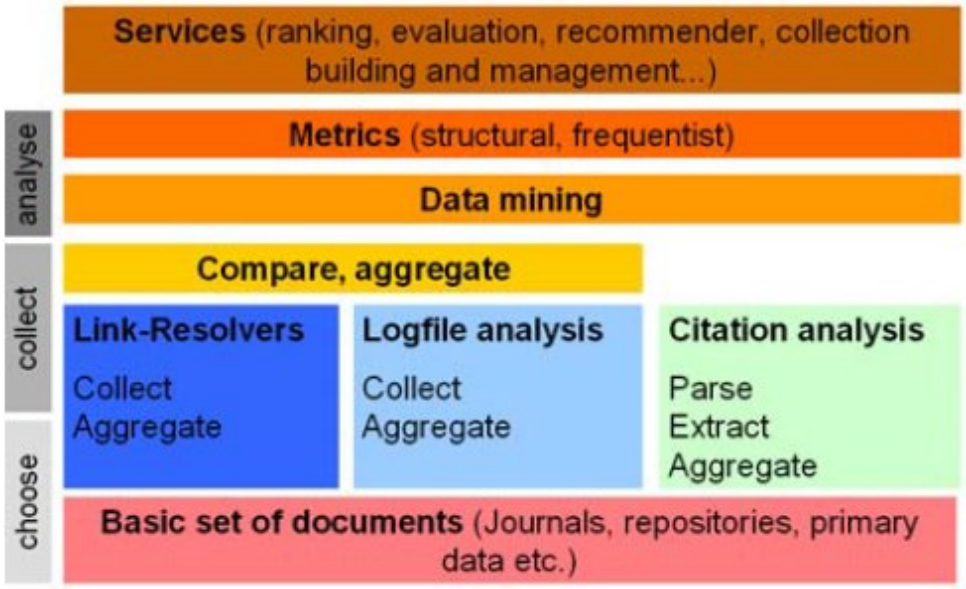
5.5 Projekty měření impaktu u publikací s otevřeným přístupem

V posledních letech vzniká několik projektů, které se snaží využít metrik stávajících či najít metriky nové pro měření impaktu vědeckých informací publikovaných v režimu Open Access, tedy volně dostupných na Internetu. Důležitost tématu dokládá např. zaměření příspěvků na konferencích ETD a IFLA, ve kterých autoři seznamují posluchače s jednotlivými projekty, vývojem a s jejich evaluací v praxi.

Vzhledem k charakteru Open Access publikování vědeckých informací se plný text práce nenachází v jednom časopise, není dostupný jen na jednom webu, ale stejný dokument může být dostupný na mnoha webových sídlech najednou a je na uživateli, jak a kde bude informace získávat. Aby bylo možné hodnotit i takovýto způsob publikování, je potřeba

a) jasně identifikovat dokument pomocí perzistentních identifikátorů, např. DOI (viz např. projekt COUNTER dále), urn:nbn, PURL, HANDLE aj. tak, aby bylo možné jednotlivé statistiky agregovat (problematika perzistentních identifikátorů přesahuje rámec této práce),
b) zajistit sběr, agregaci a zpracování jednotlivých dílčích statistických dat o využití jednotlivých dokumentů,
c) výpočet souhrnných indikátorů ze získaných dat.

Detailnější členění jednotlivých procesů navrhuje iniciativa Knowledge Exchange (80) viz Obrázek 15.



Obrázek 15 Elementy měření impaktu publikování (80)

Shrneme si proto nejprve nejdůležitější projekty zabývající se měřením impaktu publikování na Internetu a v následující části se seznámíme s projekty, které mají za cíl sběr, agregaci jednotlivých dílčích statistických dat a zpracování souhrnných indikátorů. Cílem analýzy je formulace doporučení pro měření souhrnných indikátorů pro práce distribuované v online repozitářích.

5.5.1 COUNTER

Nejznámějším standardem pro statistiky využití elektronických informačních zdrojů je standard *Counting Online Usage of Networked Electronic Resources* (zkráceně COUNTER) široce přijímaný jak producenty, tak kupujícími, knihovníky a dalšími zainteresovanými stranami. Standard definuje používané termíny, sadu metrik a reporty[26] evidující využití elektronických zdrojů – časopisů, databází, knih a multimédií v prostředí Internetu.

[26] Seznam reportů COUNTER ke stažení přes protokol SUSHI je k dispozici na http://www.niso.org/workrooms/sushi/reports/ .

Mezinárodní sada pravidel a protokolů pro zaznamenávání a sdílení dat o využití elektronických informačních zdrojů (zkráceně EIZ) je popsána v *Counter Code of Practice for eResources* (81), verze 4 je platná od roku 2012.

Reporty časopisů mj. obsahují indikátory počtu stažení HTML a PDF plných textů, což jsou nejběžnější formáty zpřístupnění EIZ. Komparace počtu stažení HTML a PDF plných textů může být zajímavým indikátorem využití EIZ – stažení plného textu v PDF může být požadováno až poté, co si vědec zobrazí náhled v HTML, potom počet stažení PDF je výrazně nižší než počet stažení HTML (82).

V poslední, čtvrté revizi *Counter Code of Practice* je pro nás významné, že byla mj. nově stanovena povinnost:

1) uvádět v reportech perzistentní identifikátor DOI pro časopisy a knihy za účelem lepší správy statistických dat a možnost propojení na online kolekce,
2) uvádět využití Open Access fulltextových článků v časopisech v samostatném reportu *Journal Report 1 Gold Open Access* (převzato z projektu SUSHI, viz dále, bohužel reporty poskytovány pouze na úrovni časopisů, nikoliv článků),
3) uvádět počet zobrazení fulltextového záznamu z výsledku, naopak povinnost statisticky vykazovat vágněji definované návštěvy (sessions) a celkové počty vyhledávání byla zrušena.

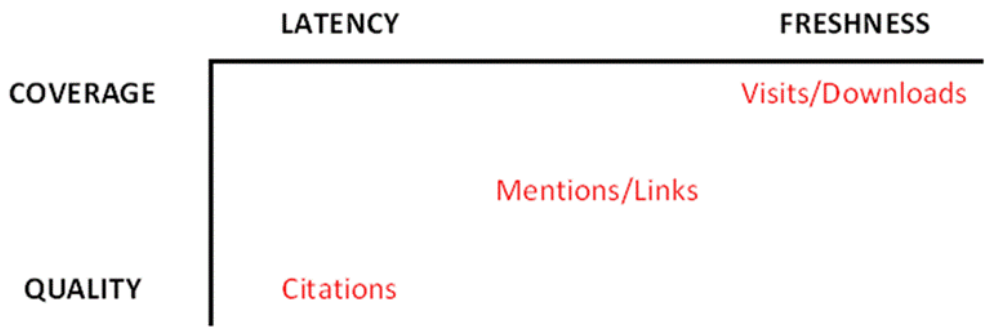
Automatizované stahování statistických dat je umožněno prostřednictvím protokolu SUSHI (viz oddíl 5.6.1 níže).

Od března 2014 jsou statistiky COUNTER měřící užití vědeckých zdrojů dostupné i na úroveň článku. *COUNTER Code of Practice for Articles* (83) definuje dvě zprávy, *Article Report 1* (počet všech úspěšných požadavků na článek, podle autora a měsíce) a *Article Report 2* (počet úspěšných požadavků na stažení plného textu článku podle autora, měsíce a DOI, sdružené podle zdrojů).

5.5.2 altmetrics

Mezi jednu z nejvýznamnějších iniciativ zaměřenou na nové metriky pro měření impaktu Open Access publikací patří projekt altmetrics, jehož autoři stanovili potřebu nových metrik ve svém programovém prohlášení *altmetrics manifesto* (84).

Nově navrhované metriky využívají (na rozdíl od tradičních metrik založených na analýze citací v tištěných vědeckých publikacích) měření událostí vyvolaných konkrétním článkem v sociálních sítích. Typicky se jedná o metriky založené na citacích/počtu odkazů na webových stránkách, počtu stažení, míře citace a aktivity na blozích a v dalších sociálních médiích (viz oddíl 5.4.6 *Sociální sítě, social bookmarking, citační manažery*). Využívané zdroje – sociální sítě – reflektují nové publikace rychleji než klasické citování v tištěných publikacích, zahrnují různé typy vědeckých výstupů i impakt na neakademickou sféru, a to v daleko větší míře nuancí než běžné metriky citační analýzy. Rozdíl mezi klasickou citační analýzou tištěných periodik a novými metrikami na bázi webometrik a sociálních sítí výstižně zobrazuje následující schéma projektu OpenAIRE (85) viz Obrázek 16.



Obrázek 16 Vliv jednotlivých metrik na pokrytí, kvalitu a reakční čas (85)

Autoři tyto nové metriky nazývají altmetrisc (v práci též použitý překlad altmetriky), neboť mohou poskytovat alternativní informace k informacím poskytovaným metrikami citační analýzy (84). Studie v oblasti altmetrik jsou prezentovány na každoročních altmetrics workshopech, jednotlivé prezentace jsou zájemcům následně většinou volně dostupné na webu projektu.

Jeden z hlavních autorů *altmetrics manifesta*, Jason Prien, zkoumá s kolegy v publikaci (86), jaké zdroje vhodných dat pro altmetriky existují a co přesněji altmetriky měří. Ze závěrů studie vyplývá, že

1) naprostá většina zdrojů altmetrik reflektuje významnou část článků (aneb s kolegy zkoumané články generují měřitelné aktivity ve zdrojích – např. citace, bookmarkování apod.),
2) aktivita v jednotlivých zdrojích se zásadně liší v čase od vydání článku i podle komunity, dáno např. odlišností chování early adopters v dané oblasti/pro daný zdroj,

3) mezi akademickými bookmarkovacími službami (Mendeley, CiteULike) a počtem citací v akademických publikacích je korelace vyšší, u obecných bookmarkovacích služeb typu Delicious korelace s citacemi prokázána nebyla,
4) je prokázána přínosnost altmetrik pro predikci počtu citací ve Web of Science.

Zásadním přínosem studie je prokázání reálnosti předpokladu využití altmetrik pro predikci počtu citací, což by mělo být předmětem dalších detailnějších výzkumů.

Společným prvkem nástrojů podle *altmetrics manifesta* je velké využití otevřených aplikačních programátorských rozhraní (zkráceně API) a souvisejících standardů (JSON, XML, HTTP/REST aj.), a to ať už pro získávání statistik ze zdrojových webů (Mendeley, Facebook, …), tak pro poskytování spočítaných altmetrik z aplikace za účelem tvorby mashupů (začlenění výsledků do vlastních aplikací). Podíváme se na dva významnější zástupce aplikací na bázi altmetrik, a to na aplikaci Impactstory pro vyhledávání vědeckých informací a výpočet altmetrics, a na Open Access časopis *PLOS ONE* prezentující altmetriky pro jednotlivé články[27].

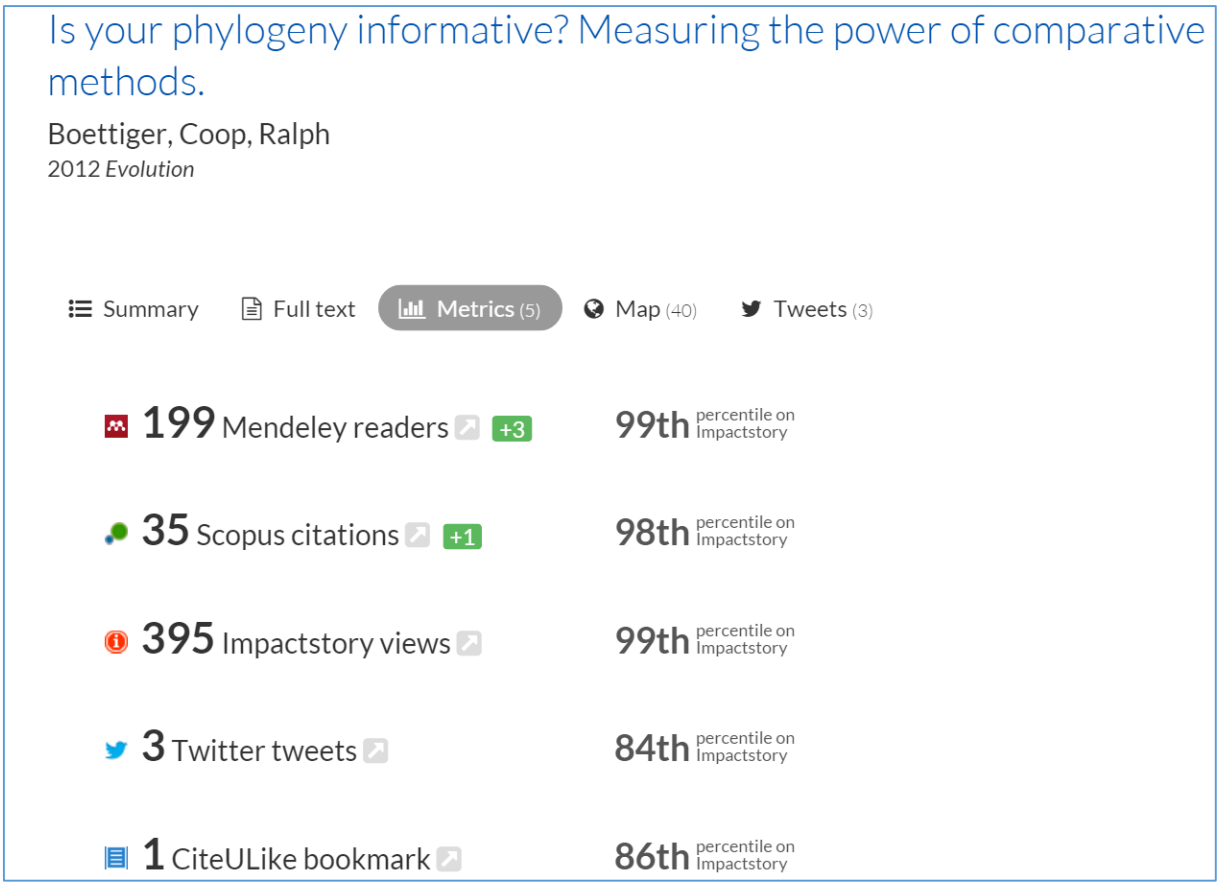
Impactstory

Webový nástroj Impactstory (http://www.impactstory.org), původním jménem Total-Impact, umožňuje uživatelskou tvorbu kolekcí dokumentů na základě různých zdrojů – seznamů DOI, účtů SlideShare apod. Pro nalezené dokumenty (články, prezentace, postery, data z výzkumů apod.) Impactstory spočítá a zobrazí metriky podle altmetrics.

Jednotlivé dokumenty v profilu autora mohou být, na základě automatického porovnání s ostatními dokumenty na Impactstory, ohodnoceny popisky typu (highly) cited, downloaded, discussed, recommended anebo viewed. Pro jednotlivé metriky (např. Mendeley readers, Scopus citations, Impactstory views, CiteULike záložky aj.) je zobrazen trend vývoje za poslední období a percentile na Impactstory pro daný rok publikace dokumentu (viz Obrázek 17).

Služba Impactstory je placená.

[27] Rozsáhlejší seznam nástrojů vznikajících na bázi projektu altmetrics lze nalézt na http://altmetrics.org/tools/



Obrázek 17 Metriky článku na webu Impactstory (120)

PLOS ONE

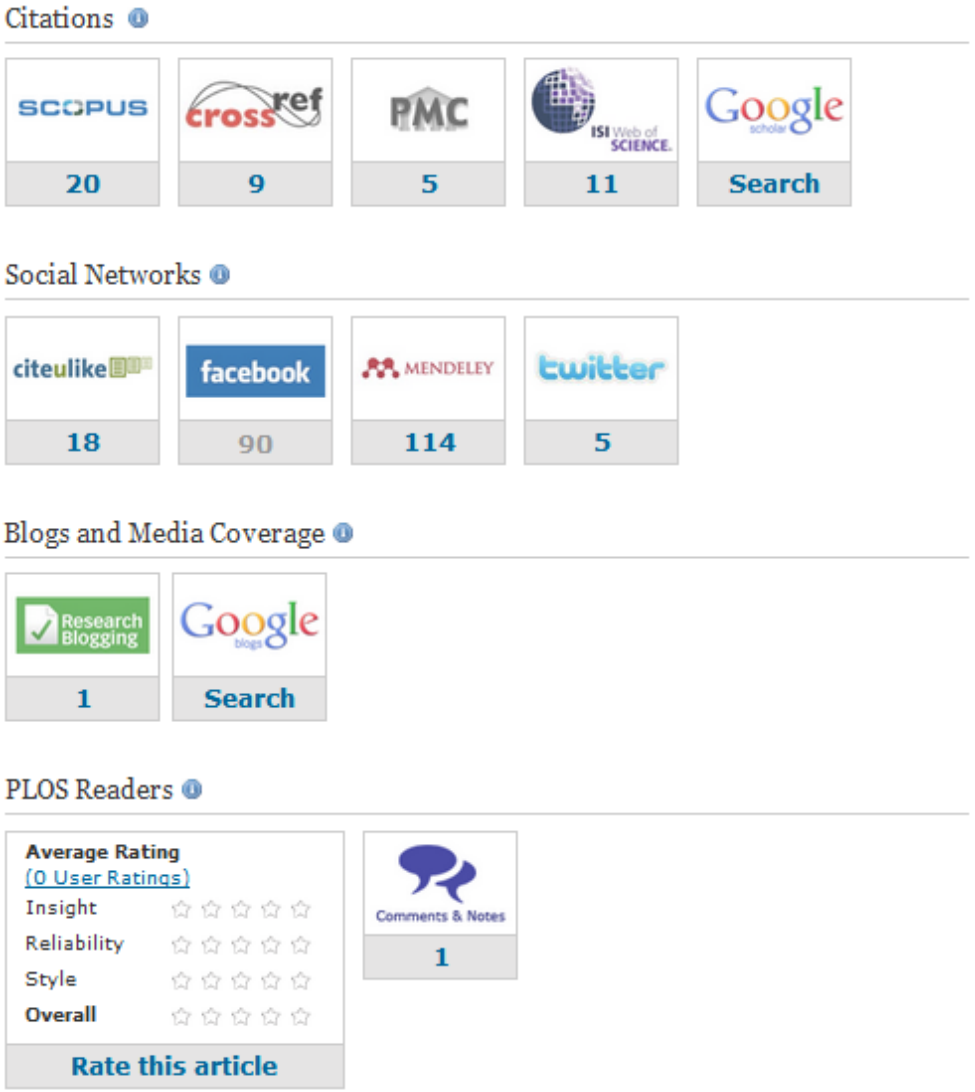
Příkladem aplikace altmetrics je časopis *PLOS ONE* (87), recenzovaný vědecký časopis zaměřený na primární výzkum, vydávaný Public Library of Science (zkratka PLOS, původně PLoS). Články z *PLOS ONE* jsou dostupné online na WWW, v režimu Open Access. Jsou indexovány významnými databázemi, jako je např. MEDLINE, PubMed Central, Scopus, Web of Science, Google Scholar aj.

Při recenzním řízení je kladen důraz na technické aspekty textu (např. preciznost provedených experimentů). Ověření důležitosti je ponecháno na veřejnosti, po zveřejnění článku. Publikované články je možné hodnotit, anotovat, komentovat, k dispozici jsou diskuse ke článkům.

Časopis *PLoS ONE* byl spuštěn v prosinci 2006 s funkcionalitou komentování a tvorby poznámek. Později přibyly funkce hodnocení článků, zpětných odkazů (trackbacks). Vydavatel PLoS začal v souladu s projektem Article Level Metrics zveřejňovat online

uživatelská data pro publikované články, v současnosti *PLOS ONE* poskytuje řadu indikátorů včetně citačních metrik, statistik využití, pokrytí v blozích, social bookmarking, komunitního a expertního hodnocení (viz Obrázek 14 a Obrázek 18). Využití jednotlivých indikátorů pro zhodnocení kvality a impaktu článku na základě jednotlivých metrik je již ponecháno na čtenáři.

Poplatek za vydání článku ve výši 1 350 USD – 2 900 USD (podle časopisu) v některém z Open Access časopisů vydavatelství PLOS platí autor, články vychází pod licencí Creative Commons Attribution License (CC-BY).



Obrázek 18 Metriky PLOS ONE založené na citační analýze, sociálních sítích a hodnocení čtenářů (87)

5.6 Agregace a zpracování statistických dat

V případě metrik na bázi citační analýzy tištěných periodik jsou data zadávána do specializované databáze experty (vyšší přesnost, menší pokrytí zdrojových časopisů) nebo automaticky získávána extrakcí textu a metodami data miningu (s nižší přesností, ale s vyšším pokrytím).

Pro výpočet metrik u eVŠKP a dalších dokumentů publikovaných v prostředí Internetu je zapotřebí zajistit přístup ke statistickým podkladům ve vhodné formě, a to nejlépe pomocí dobře zdokumentovaného API a webových standardů. Potřeba se týká především článků publikovaných v jiných jazycích než v angličtině a v rozvojových zemích (kde je pravděpodobnost citovanosti v impaktovaném časopise menší), či článků v Open Access repozitářích, které nejsou publikovány zároveň v Open Access impaktovaném časopisu a nejsou tak zahrnuty v relevantních databázích typu WoK, JCR, Scopus apod. používaných pro výpočet JIF a h-indexu (75).

Pro indikátory typy altmetrics se využívá API jednotlivých zdrojových databází/aplikací, většinou se neuplatňuje agregace dat z jednotlivých odlišných zdrojů pro výpočet jednoho uceleného indikátoru, ale prezentuje se více indikátorů. Pro výpočet webometrických indikátorů na bázi využití se používá především logů webových serverů o využití konkrétního časopisu či dokumentu (např. záznamy stažení WWW stránek či PDF), kde agregace dat z více zdrojů (repozitářů) se stejným dokumentem má naopak velký význam. V posledních letech se objevuje několik projektů zaměřených na sběr a zpracování takovýchto statistických dat, které autor představuje v této podkapitole. Jak bude vidět, projekty jsou úzce provázané, autoři vzájemně spolupracují a doplňují se.

5.6.1 SUSHI

Standardizovaný protokol SUSHI (*Standardized Usage Statistics Harvesting Initiative*) ANSI/NISO Z39.93-2007 (82) definuje model automatizovaných dotazů a odpovědí (webových služeb SOAP) pro harvestování statistických výkazů COUNTER (či jim obdobných) o využití elektronických informačních zdrojů.

Knihovníci nemusí data stahovat ručně či nechat si je zasílat e-mailem, naopak díky standardizaci mohou začlenit automatické stahování do systémů na správu elektronických informačních zdrojů. Data jsou stahována ve formátu XML.

Protokol SUSHI sám o sobě neřeší agregaci statistických dat pro stejný zdroj v různých databázích, ale díky nové povinnosti uvádět DOI ve statistikách COUNTER by toto mohl řešit systém správy EIZ.

5.6.2 PIRUS, PIRUS2

Program *Publisher and Institutional Repository Statistics* (zkráceně PIRUS) z roku 2008 sponzorovaný JISC měl za úkol vytvořit a upravit standardy, statistiky a procesy za účelem rozšíření standardu COUNTER na úroveň časopiseckých článků. Projekt demonstroval možnost tvorby, zaznamenávání a agregace takovýchto statistik využití, navržené XML schéma pro statistiky využití na úrovni jednotlivých článků bylo implementováno např. provozovateli repozitářů PLOS a SURF.

Navazující program PIRUS2 probíhal od listopadu 2009 do prosince 2010, jeho cílem bylo mj. ověřit škálovatelnost navržených metod, stanovit náklady na tvorbu reportů a centralizovanou správu a rozšířit akceptaci navržených metod mezi další provozovatele repozitářů, vydavatele a autory.

Projekt PIRUS2 doporučuje (88) identifikovat zdrojové články pomocí metadatového balíčku OpenURL ContextObject pro časopisecké články[28]. V rámci projektu byly připraveny moduly pro repozitáře na bázi DSpace, EPrints a Fedora pro zpřístupnění dat o využití časopiseckých článků. Implementaci takto ověřené možnosti zpřístupňování a zpracování statistik na úrovni článků do protokolu COUNTER projekt doporučuje k dalšímu výzkumu.

Projekt PIRUS2 navrhl rozšíření projektu COUNTER o report *Article Report 1* – počet úspěšných požadavků na plný text článků podle DOI a měsíce, ve formátu Excel (ruční stažení) a XML (automatické zpracování, protokol SUSHI). Jako hlavní způsob konsolidace dat od jednotlivých poskytovatelů je navrženo využití perzistentního identifikátoru DOI.

[28] Bližší popis metadatových prvků viz http://ocoins.info/cobg.html.

Projekt PIRUS navrhl ustanovení centrální clearingové instituce (tzv. Clearing House) zajišťující sběr, konsolidaci dat a distribuci reportů o využití a navrhl ekonomický model provozu. Při události stažení plného textu je dle projektu přiřazen události kód (tzv. tracker code) a vygenerován OpenURL záznam v logu. Projekt PIRUS původně navrhoval 3 metody následného zpracování logů:

A. OpenURL záznam je zaslán na lokální server, kde je filtrován na základě pravidel COUNTER (eliminování robotů a tzv. dvojkliků), vygenerovány statistky využití COUNTER pro článek identifikovaný DOI v XML formátu a jejich následné zpřístupnění (vhodné pro velké repozitáře a producenty, preferováno)
B. OpenURL záznam je uložen na lokální OAI-PMH server, odkud je stažen protokolem OAI-PMH do Clearing House, kde je provedeno zpracování podle pravidel COUNTER a vygenerování statistik obdobně jako v případě lokálního serveru ve scénáři A, statistiky jsou následně zpřístupněny autorizovaným institucím.
C. OpenURL záznam je zaslán do Clearing House, kde je provedeno zpracování podle pravidel COUNTER a vygenerování statistik obdobně jako v případě lokálního serveru ve scénáři A, statistiky jsou následně zpřístupněny autorizovaným institucím.

Oproti návrhu projektu PIRUS scénáře B na využití protokolu OAI-PMH pro harvestování statistických dat (pro srovnání – OAI-PMH používají projekty Open Access Statistics a KE/SURFsure viz níže), projekt PIRUS2 klade důraz na zbývající dvě navržené metody zpracování dat a možnost vystavení dat pomocí OAI-PMH ponechává jako doplňující. (88)

V případě užití scénářů B a C při předávání a zpracování podkladových statistických dat je jednodušší agregace dat z různých repozitářů, využitím centrálního zpracování na základě pravidel COUNTER je zaručen konzistentní způsob zpracování nad všemi zapojenými repozitáři. Vzhledem k nižším nárokům na implementaci na straně lokálních repozitářů by implementace měla být jednodušší. (89)

5.6.3 Open Access Statistics

Problém chybějící standardizace v oblasti metrik využití vědeckých článků v prostředí Internetu se rozhodli řešit autoři, především z řad německých knihoven, projektu Open Access Statistics (OAS). Projekt vychází ze standardů COUNTER, LogEc (agregace statistik

přístupů služeb RePEc, na úrovni článků) a IFABC (metriky pro WWW servery, dokumenty, e-maily aj.).

Příkladem problémů řešených v projektu je přenos podkladových dat/logů a deduplikace uživatelů/dokumentů nejen v rámci jednoho repozitáře, ale také v síti Open Access repozitářů, kdy jeden uživatel může volně přecházet mezi různými repozitáři s identickými verzemi dokumentů. Centrální zpracování dat dle projektu umožňuje také sledovat trend stahování různých dokumentů jedním uživatelem z různých repozitářů, příp. součet požadavků na různé dokumenty příbuzného zaměření v odlišných repozitářích.

V rámci projektu OAS byla v první fázi (květen 2008 – prosinec 2010) vybudována síť repozitářů za účelem sběru a výměny informací o využití jednotlivých časopiseckých článků.

Role poskytovatele dat (75):

- generování logů o využití dokumentů,
- pseudo-anonymizace citlivých informací, např. IP adres,
- zpracování informací o využití (přidání ID dokumentu, metadatový popis OpenURL ContextObject),
- vystavení informací prostřednictvím OAI-PMH serveru pro stažení centrální službou.

Role centrálního serveru:

- harvestování informací od jednotlivých poskytovatelů,
- deduplikace dokumentů (např. identické dokumenty z odlišných repozitářů),
- deduplikace uživatelů,
- zpracování dat dle standardů COUNTER, LogEc a IFABC (odstranění automatizovaných přístupů např. roboty, opakované klikání/zobrazování i mezi spolupracujícími servery apod.).

V druhé fázi (duben 2011 – duben 2013) se projekt OAS, resp. OAS2 zaměřil na standardizaci a ověření indikátorů na bázi absolutní frekvence využití dokumentů, standardizaci procesů, uložení dat, rozhraní pro výměnu dat o využití, zajištění trvalé udržitelnosti a na integraci nových služeb, jako je např. řazení na bázi frekvence stahování, indikace impaktu dokumentu aj.

5.6.4 KE Usage Statistics Group, SURFsure

Iniciativa Knowledge Exchange (KE) má za cíl podporovat spolupráci mezi národními institucemi v Evropě odpovědnými za vývoj infrastruktury a služeb na podporu ICT ve vědě a výzkumu. V rámci aktivit zaměřených na spolupráci digitálních repozitářů, konkrétně na jejich statistiky využití, uspořádala řadu seminářů, kterých se účastnili odborní zástupci souvisejících projektů COUNTER, PIRUS, OAS a SURF Statistics on the Usage of Repositories (SURFsure, projekt na duben 2009 – březen 2010), jakož i např. zástupci služeb RePec a NeeO zaměřených na Open Access v ekonomii.

V březnu 2010 v rámci semináře ke statistikám využití v Berlíně byla iniciována příprava vize statistik využití, na jejímž základě byl zhotoven dokument *Combined usage statistics as a basis for Research intelligence* (80). V dokumentu autoři argumentují pro sběr a výměnu statistik o využití, které nabízejí cenné informace nejen pro vědeckou komunitu, ale také pro obchod a společnost jako celek. Získané důvěryhodné indikátory o Open Access publikacích mohou posloužit také jako podklad pro informované rozhodování v oblasti vysokoškolského vzdělávání a výzkumu.

Vzhledem k absenci jednotného standardu sdílení statistik o využití mezi jednotlivými systémy a s tím spojených problémů, byl v rámci pracovní skupiny KE připraven k podzimu 2010 dokument *KE Usage Statistics Guidelines* (89), který vycházel z již probíhajících prací v souvisejících projektech. Tyto pokyny jsou nyní součástí projektu SURFsure.

Využití dokumentu je definováno jako zobrazení plného textu dokumentu nebo jemu přidružených metadat. Podobně jako je definováno v projektu PIRUS ve scénáři B (a využito např. v rámci OAS), pro stažení statistických dat clearingovým centrem je využito protokolu OAI-PMH, nebo je možné užít protokol SUSHI viz scénář C projektu PIRUS.

V rámci *KE Usage Statistics Guidelines* je kromě způsobu přenosu také řešeno:

a) jednotný formát přenášených dat na bázi OpenURL Context Object, jehož vhodnost byla prokázána v předešlých projektech,

b) normalizace dat – filtrování „double clicks“ v clearingovém centru (opakované požadavky jedním uživatelem) a filtrování přístupů robotů v lokálních repozitářích a příp. pokročilejší heuristikou v clearingovém centru (seznam robotů je vytvořen kombinací seznamů z projektů COUNTER, AWStats, Universidade do Minho a PLOS),
c) pseudo-anonymizace dat – IP adres v souladu se *Směrnicí 95/46/ES o ochraně fyzických osob v souvislosti se zpracováním osobních údajů a o volném pohybu těchto údajů*.

Na základě výše uvedených pokynů jsou zpracována mj. Pravidla OpenAIRE určená pro získávání statistik využití ze zapojených Open Access repozitářů v rámci pilotního projektu otevřeného přístupu v 7. rámcovém programu EK (85).

5.6.5 Projekt IRUS-UK

Za finanční podpory JISC byl vybudován projekt IRUS-UK (90), který vychází z doporučení stanovených v projektech PIRUS (viz oddíl 5.6.2). V rámci projektu byla vybudována centrální clearingová instituce (tzv. Clearing House) zajišťující sběr, konsolidaci a distribuci reportů podle statistik Counter (viz oddíl 5.5.1) o počtu stažení záznamů z institucionálních repozitářů Velké Británie. Projekt byl prezentován Paulem Needhamem na konferenci ETD 2014 v Leicesteru, UK (91).

IRUS-UK zajišťuje sběr základních dat ze spolupracujících repozitářů, která zpracovává do statistik odpovídajících standardu Counter, resp. vychází z doporučení *COUNTER Code of Practice* (81) a *PIRUS Code of Practice* (92). Poskytuje tak srovnatelné, autorizované a standardizované údaje o využití napříč repozitáři ve Velké Británii.

Konkrétní postup implementace pro IRUS-UK je popsán v dokumentu *IRUS-UK Code of Practice*, který definuje technické, organizační a ekonomické modely pro záznam, vykazování a sběr statistik využití všech typů záznamů zpřístupněných ve spolupracujících institucionálních repozitářích ve Velké Británii.

Sběr dat z institucionálních repozitářů probíhá podle specifikace *The Tracker protocol V3.1* (93). Podle této specifikace, pokud uživatel klikne na odkaz pro stažení souboru z repozitáře, je odeslán na vzdálený server k dalšímu zpracování OpenURL záznam ve formátu

ContextObject (podle NISO standardu OpenURL 1.0). OpenURL řetězec obsahuje údaje ve formátu klíč=hodnota oddělené znakem & (viz Tabulka 2).

Tabulka 2 Prvky IRUS-UK Push protokolu (zdroj: vlastní zpracování podle (93))

Prvek	OpenURL klíč	PZopis
OpenURL version	url_ver	Identifikace dat podle OpenURL 1.0
Usage event datestamp	url_tim	Časové razítko události (datum a čas)
Client IP address	req_id	IP adresa klienta požadujícího článek
UserAgent	req_dat	Řetězec UserAgent identifikující klientský program
Item OAI identifier	rtf.artnum	OAI identifikátor
FileURL	svc_dat	URL na plný text
HTTP Referer	rfr_dat	Pole HTTP Referer podle HTTP protokolu
Source repository	rfr_id	Identifikátor repozitáře, ve kterém došlo k události

Následně jsou záznamy v centrálním registru deduplikovány podle pravidel 5. sekce 4. revize *COUNTER Code of Practice* (81) o záznamy generované roboty (seznam robotů podle Counter doplněný o vlastní seznam IRUS-UK) a záznamy generované neobvyklým stahováním dat (dvojkliky a většinou více jak 100 stažení z repozitáře z jedné IP adresy denně).

Na základě zpracovaných dat je generována řada standardizovaných statistik, které jsou dostupné zapojeným institucím na webu projektu IRUS-UK.

5.7 Evaluace vhodnosti altmetrik pro eVŠKP

Ukazuje se, že klasické citační metriky neposkytují dostatečné podklady pro evaluaci publikační činnosti v režimu Open Access, na úrovni jednotlivých článků. Nově navrhované metriky jako např. altmetrics obsahují řadu nových indikátorů, ale neposkytují jeden ucelený kvantifikovatelný indikátor. Výběr vhodných indikátorů a jejich váha je tak na uživateli a jeho konkrétních potřebách.

Naskýtá se otázka, zda altmetriky jsou vhodné pro eVŠKP, u kterých lze očekávat – pro jejich nižší vědecký význam – nižší míru citování a zmiňování na sociálních sítích. Ověřme proto hypotézu, že míra odezvy eVŠKP na sociálních sítích je málo významná, tj. počet prací z testovaného repozitáře s ohlasem vyšším jak 1 je v dané sociální síti[29] pod 1 %. Míra odezvy vyšší než 1 je stanovena pro eliminaci potenciálních autocitací, kdy sám autor v sociální síti odkazuje na svoji vlastní práci.

Pro ověření hypotézy autor disertační práce využil statistik služby PlumX. „PlumX™ je nástroj pro impakt, který poskytuje přehled o využití výsledků vědy ve všech jejích formách, o interakci s výsledky a mluvení o výsledcích po celém světě. Díky sklízení metrik z mnoha online zdrojů nabízí PlumX vzrušující nový pohled na vliv výzkumníků, skupin a institucí. Pro získání přehledu již nemusíte čekat roky, až je výzkum případně citován. Neomezujte se jen na články – PlumX sleduje desítky typů výstupů. Díky souhrnům a vizualizacím dat můžete rychle vidět metriky, které předtím nebyly k dispozici. Můžete se také do dat ponořit za využití nástrojů PlumX nebo exportovat data do nástrojů, které právě používáte.” (94)

Pro testování autor zvolil repozitář Pittsburské univerzity, konkrétně část týkající se eVŠKP sledovanou v rámci nástroje PlumX. Z webu PlumX byl dne 25. 7. 2014 získán CSV export s nalezenými ohlasy v níže uvedených sledovaných nástrojích.

[29] Jedna zmínka v sociální síti, např. sdílení na Facebooku, bude pravděpodobně vytvořena samotným autorem, a proto v naší hypotéze zvažujeme pouze hodnoty vyšší.

Sociální sítě:

- Delicious
- Google+
- Facebook
- Twitter

Stažení:

- Bitly - kliknutí na zkrácené URL
- repozitář Pittsburské univerzity – počty stažení plného textu

Po pročištění a importu dat do Excelu autor získal údaje 5 307 prací. Zpracované souhrnné statistiky zobrazuje Tabulka 3.

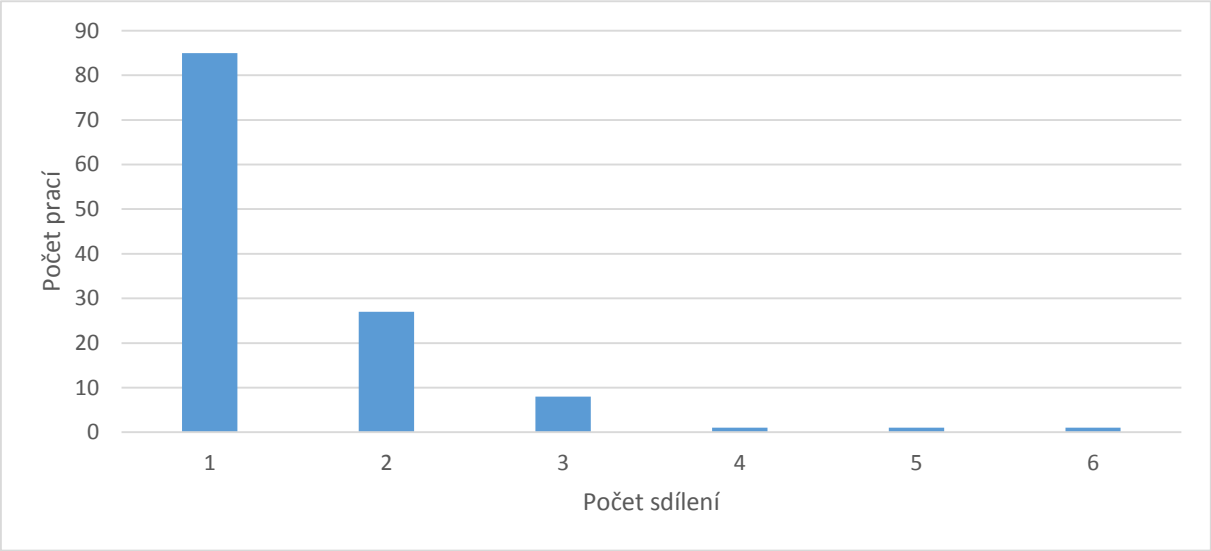
Tabulka 3 Statistické vyhodnocení metrik PlumX Pittsburské univerzity (zdroj: vlastní zpracování)

	Delicious	Google+	FB Shares	FB Likes	Tweets	FB Comments	Pitts Downloads	Bitly Clicks
Zmínky	1	14	178	302	44	178	1 070 540	83
Zmínky / celkem	0,00	0,00	0,03	0,06	0,01	0,03	201,72	0,02
Nálezů	1	12	123	38	26	31	5013	15
Nálezů / celkem	0,02 %	0,23 %	2,32 %	0,72 %	0,49 %	0,58 %	94,46 %	0,28 %
Zmínky / nálezů	1,00	1,17	1,45	7,95	1,69	5,74	213,55	5,53

Řádek „Zmínky“ zobrazuje celkový počet všech nalezených zmínek o sledovaných kvalifikačních pracích v konkrétní službě. Průměrný počet zmínek jedné eVŠKP, údaj „Zmínky / celkem“, je vypočten jako podíl všech nalezených zmínek k počtu sledovaných záznamů, tj. 5 307. Jako „Nález“ je započítána taková kvalifikační práce, o které je evidována min. jedna zmínka v dané službě. Poslední dva řádky nám zobrazují procentuální podíl eVŠKP, u kterých byla nalezena zmínka v konkrétní službě („Nálezů / celkem“) a průměrný počet zmínek u prací s pozitivním nálezem („Zmínky / nález“; tj. zde pomíjíme práce, u kterých v databázi PlumX není evidován záznam o využití v konkrétním nástroji).

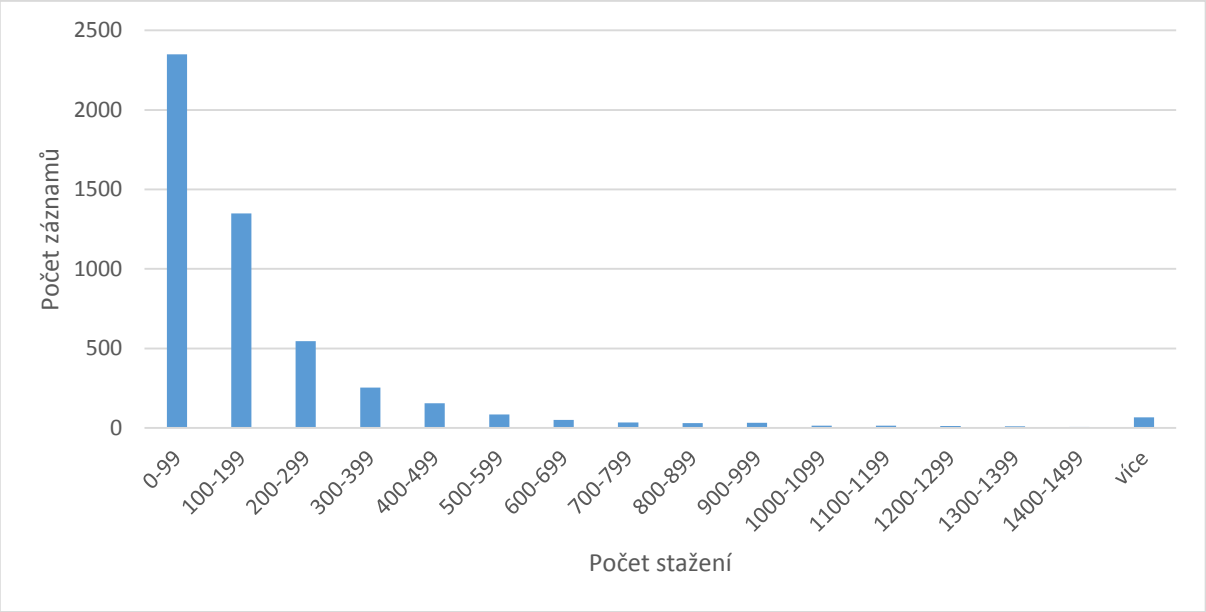
Podle výsledků webu PlumX jsou nad daty repozitáře kvalifikačních prací Pittsburské univerzity nejméně využívané tyto služby: nástroj Delicious pro ukládání, sdílení a objevování URL záložek (1 využití), sociální síť Google+ (14 využití), Twitter (44 tweetů) a nástroj pro zkracování URL adres Bitly (83 kliknutí na zkrácenou URL).

Ze sociálních sítí je nejvyužívanější Facebook, pro který je evidováno 178 sdílení u 123 prací. Jedno sdílení má v průměru 1,7 označení Like a 1 komentář. Pravděpodobnost, že práce z repozitáře Pittsburské univerzity bude alespoň jednou sdílena na Facebooku je 2,32 %. Pokud však eliminujeme záznamy sdílené na síti Facebook právě jednou, jak je stanoveno ve zkoumané hypotéze, získáme pouhých 38 nálezů namísto původních 178, což odpovídá 0,01 % záznamů v repozitáři. Detailnější rozložení počtu sdílení na Facebooku u nálezů znázorňuje Graf 18. Můžeme proto předpokládat, že naprostou většinu záznamů sdílí na Facebooku pouze sami autoři, takovýto nález však pro nás nemá vypovídající schopnost o významu kvalifikační práce.



Graf 18 Histogram počtu sdílení prací na Facebooku podle PlumX (zdroj: autor)

Poslední statistika sledovaná v nástroji PlumX je počet stažení plného textu z institucionálního repozitáře univerzity. Histogram rozdělení počtu záznamů v repozitáři univerzity podle počtu stažení znázorňuje Graf 19.



Graf 19 Histogram počtu stažení prací z repozitáře podle PlumX (zdroj: autor)

Výše uvedené výsledky altmetrik s podílem nálezů pod 1 % potvrzují, nad daty z repozitáře kvalifikačních prací Pittburské univerzity, pro všechny analyzované služby platnost hypotézy o nízké míře využití altmetrik mezi uživateli repozitáře eVŠKP. Počty stažení jsou však již statisticky významné a mohou posloužit jako jedna z metrik indikující významné eVŠKP v repozitáři.

5.8 Závěr kapitoly

V případě rozsáhlých Open Access repozitářů s významnými výsledky vědy a výzkumu nebo v případě publikace novinových článků na aktuální a nová témata se jako zajímavou alternativou klasických citačních metrik jeví altmetriky. Podmínkou využití je atraktivnost dokumentu a jeho časté sdílení v sociálních sítích komunitou čtenářů. Příkladem vhodného využití altmetrik je Open Access časopis *PLOS ONE* (viz oddíl 5.5.2).

U méně významných prací, jako jsou vysokoškolské kvalifikační práce, se však nepodařilo prokázat přínos altmetrik pro hodnocení významu práce z důvodu nižší míry citovanosti prací na sociálních sítích (viz podkapitola 5.7), vyplývající z nižšího významu a menší atraktivnosti eVŠKP pro internetovou komunitu. Vhodné metriky pro repozitáře eVŠKP by proto měly být založeny spíše na webometrikách, jako je např. počet stažení plného textu (viz oddíl 5.4.4).

U eVŠKP mohou být identické plné texty dostupné v několika repozitářích najednou, např. v ČR v repozitáři univerzity, v registru Theses.cz a v Národním úložišti šedé literatury. Je proto potřeba zajistit jednotnou agregaci, deduplikaci a interpretaci statistik z individuálních repozitářů. V této oblasti je důležitá mj. iniciativa Knowledge Exchange (viz oddíl 5.6.4), která zastřešuje jednotlivé obdobné projekty včetně projektů COUNTER a PIRUS. Jako vhodný formát přenosu statistických dat doporučuje využít v repozitářích protokolů SUSHI nebo OAI-PMH.

V případě Open Access repozitářů je vhodné sbírat statistické podklady pro výpočet webometrických indikátorů a vypočítané metriky nabídnout uživatelům tak, aby se usnadnilo jejich rozhodování o důležitosti a vhodnosti jednotlivých článků v repozitáři. Příkladem takovéto prezentace dat může být např. zobrazení nejstahovanějších prací v daném období, viz např. DART-Europe E-theses Portal popsaný v oddílu 2.3.1.

Pro zpracování statistik nad repozitáři eVŠKP v České republice lze doporučit vybudování centrálního agregačního bodu, který by sbíral a jednotně vyhodnocoval statistiky pomocí jednotlivých metod a standardů. Vzorová implementace vhodná i pro podmínky repozitářů ČR byla popsána na projektu IRUS-UK v oddílu 5.6.5. Prozatím takováto služba v ČR neexistuje, míru užití eVŠKP je proto možné vyhodnocovat pouze na úrovni jednoho repozitáře.

V roce 2015 Jan Mach vede bakalářskou práci na téma agregace metrik užití.

6 Výběr systému centralizovaného vyhledávání

Jednou ze základních částí procesu zpřístupňování eVŠKP je vyhledávání záznamů v repozitářích. Metadata, příp. plné texty eVŠKP mohou být vyhledávány jak v samotném primárním repozitáři, tak i v katalogu knihovny nebo ve službě centralizovaného vyhledávání nad jednotným indexem agregátora záznamů.

Výhodou vyhledávacího rozhraní repozitáře je možnost přesné formulace dotazu za využití široké škály evidovaných metadat. Vyhledávání v konkrétním repozitáři využijeme především tehdy, pokud máme přesnější představu o hledané eVŠKP (známe např. fakultu či katedru, školitele, rok obhajoby, získaný titul apod. selekční údaje specifické pro eVŠKP). Řešení tvorby indexu z metadat ve formátu EVSKP-MS a vyhledávacího rozhraní nad repozitářem eVŠKP je věnována kapitola následující.

Pokud uživatelé nemají konkrétní představu o požadovaném typu dokumentu, znají jen tematické zaměření a odpovídající klíčová slova, je možné využít služeb agregátora záznamů s centrálním indexem, který obsahuje metadata a příp. plné texty ze širokého spektra zdrojů poskytovaných uživatelům univerzitou. Jedná se konkrétně o placené elektronické informační zdroje, online zdroje dostupné v režimu open access, ale i záznamy univerzity – ze souborného katalogu, z repozitáře eVŠKP a z případného institucionálního repozitáře vědeckých prací. Provozovatel centralizovaného vyhledávání má za úkol zajistit pravidelné nabírání metadat a plných textů od jednotlivých poskytovatelů dat, jejich indexování, deduplikaci a vyhledávání nad centrálním indexem.

„Google generace“ uživatelů upřednostňuje služby fungující na principu internetových vyhledávačů typu Google, s jedním vyhledávacím polem, bez znalosti specifických metadat a klasifikací, s výsledky řazenými podle relevance. Preferují prohledávání všech zdrojů z jednoho místa, bez nutnosti ručně kombinovat záznamy nalezené v odlišných rozhraních jednotlivých poskytovatelů služeb. Oproti systémům federativního vyhledávání, discovery služby poskytují rychlou odezvu na dotaz, nejsou závislé na funkčnosti jednotlivých dílčích vyhledávacích služeb, umožňují deduplikaci záznamů, obohacování metadat jejich kombinováním z různých zdrojů, využití faset pro zpřesňování obecně položených dotazů.

Služba centralizovaného vyhledávání je také nazývána pojmem *discovery služba*, neboť umožňuje uživatelům „objevovat“ záznamy z elektronických informačních zdrojů včetně

repozitářů eVŠKP, které by při vyhledávání v pouze několika osobně preferovaných zdrojích neprohledávali.

Záznamy z repozitáře eVŠKP univerzity jsou uživatelům snadno dostupné v rámci jednotného vyhledávacího rozhraní knihovny díky discovery službě, která musí podporovat import metadat z lokálních repozitářů univerzity do centrálního indexu (typicky protokolem OAI-PMH, případová studie implementace viz podkapitola 4.3). Cílem této kapitoly je poskytnout doporučení, jak správně formulovat zadávací dokumentaci pro výběr discovery služby a vyhnout se tak potenciálním problémům, které mohou nastat.

Na základě žádosti ředitele knihovny Univerzity Jana Evangelisty Purkyně v Ústí nad Labem, Ivo Brožka, autor disertační práce v roce 2013 vypracoval podklady pro zadávací dokumentaci k veřejné zakázce *Systém centralizovaného vyhledávání elektronických informačních zdrojů* (4) a zúčastnil se hodnocení nabídek jako přizvaný specialista, formulovaná doporučení pro výběr discovery služby jsou uvedena níže.

Při analýze realizovaných výběrových řízení v ČR na discovery služby (např. na Univerzitě Pardubice (95, 96) a v Národní technické knihovně (97)) byly identifikovány kritické faktory úspěšnosti výběrových řízení, které byly vzaty v potaz při přípravě doporučení. Jedná se o následující rizika:

- špatná specifikace požadovaných indexovaných zdrojů může zapříčinit výhru discovery služby, která klíčové zdroje pro univerzitu neindexuje,
- nízká transparentnost míry indexace zdrojů – např. producent indexuje pouze část z daného časopisu, indexace neúplných dat z alternativních zdrojů, uvádění pouze souhrnného procenta indexace bez specifikace konkrétních publikací a s tím související netransparentní výpočet míry pokrytí uchazečem, nejasná specifikace požadovaných údajů metadat apod.,
- výběr discovery služby, která neposkytuje klíčové služby a technickou podporu požadované univerzitou,

- nevhodně stanovené váhy subkritérií, např.:
    - Pokud je zadavatelem stanovena vysoká váha subkritéria cena za roční přístup a nízká váha subkritéria cena instalace, může výběrové řízení vyhrát nabídka s nejvyšší cenou za instalaci systému, ale jen 1 Kč za roční přístup ke službě – – tj. nabídka s nejvyšší celkovou cenou nabídky.
    - Může vyhrát nabídka sice s nízkou cenou, ale s plně nevyhovující funkčností.

Doporučení výběru discovery služby v této kapitole byla vypracována také na základě rozsáhlých konzultací autora se zástupcem zadavatele a se zástupci discovery služeb ExLibris PRIMO, ProQuest Summon a EBSCO Discovery Service.

Níže jsou uvedeny klíčové části zadávací dokumentace, založené na autorově analýze, které jsou důležité pro eliminaci uvedených rizik.

Údaje specifické pro konkrétního zadavatele jsou v níže uvedených odstavcích nahrazeny popisem požadovaného údaje v hranatých závorkách, např. [počet].

6.1 Druh zadávacího řízení

Podlimitní veřejná zakázka na služby.[30]

Otevřené řízení dle § 27 zákona č. 137/2006 Sb., o veřejných zakázkách, ve znění pozdějších předpisů.

6.2 Předmět veřejné zakázky

Klasifikace předmětu veřejné zakázky (NIPEZ kódy):

- 48160000-7 – Balíky programů pro knihovny (Hlavní předmět)
- 72414000-5 – Poskytovatelé webových vyhledávačů
- 71356300-1 – Technická podpora

[30] Za veřejnou zakázku malého rozsahu se považuje taková veřejná zakázka, jejíž předpokládaná hodnota nedosáhne v případě dodávek a služeb 2 000 000 Kč. Jednotlivé školy si stanovují většinou vlastní limit hodnoty zakázky (např. VŠE v Praze a Univerzita Karlova 250 000 Kč), při jehož překročení je škola povinna vybrat dodavatele výběrovým řízením, např. uzavřenou nebo otevřenou výzvou, kdy je právě vhodné využít doporučení z této kapitoly.

Předmětem plnění veřejné zakázky je poskytnutí služby centralizovaného vyhledávání elektronických informačních zdrojů (discovery systému) ve formě Software as a Service, link serveru na bázi OpenURL pro potřeby odkazování na nalezené záznamy dostupné v EIZ a související potřebné softwarové licence, konfigurace služeb dle požadavků Zadavatele a zajištění dostupnosti služeb po smluvně dané období.

Součástí veřejné zakázky je také technická a provozní podpora, tj. zajištění provozu, údržby, drobných úprav, update, upgrade, aktualizace indexu novými metadaty, rozvoj vyhledávacího systému včetně přidávání dalších zdrojů EIZ, zajištění služeb helpdesk a provádění servisních zásahů.

Předmětem plnění veřejné zakázky jsou zejména tyto činnosti:

1. Jednotné vyhledávací rozhraní pro elektronické informační zdroje, záznamy katalogu knihovny Zadavatele a repozitáře Zadavatele [název repozitáře eVŠKP].
2. Link server, A-Z seznam časopisů.
3. Instalace na serverech poskytovatele služby, konfigurace systému.
4. Průběžná aktualizace systému a dat.
5. Provoz na serverech poskytovatele služby, systémová podpora.
6. Související licence.

6.3 Minimální technické parametry

Následující výčet uvádí minimální technické parametry dodávaného systému, které musí být bezpodmínečně splněny v rámci předmětu této veřejné zakázky. Nabídky, které nesplňují některé z níže uvedených kritérií, budou ze zadávacího řízení vyloučeny z důvodu nesplnění zadávacích podmínek.

1. Technické řešení je založeno na předem vytvořeném centrálním indexu pro všechny prohledávané EIZ, vytvořeném na smluvním základě s jednotlivými vydavateli a producenty dat. Centrální index pravidelně indexuje min 60 % metadat (minimálně název, autor, rok vydání, zdroj – název titulu, ročník, číslo pro časopis) a min. 60 % plných textů dostupných v licencovaných EIZ Zadavatele podle Tabulky *Pokrytí indexu* [31].

[31] Tabulka *Pokrytí indexu* Zadávací dokumentace obsahuje seznam poptávaných elektronických informačních zdrojů, u každého zdroje je stanovena $váha_1$ a $váha_2$ viz Bodové hodnocení níže.

2. Discovery systém umožňuje vyhledávání záznamů a následné zúžení výsledkové množiny záznamů pomocí faset podle různých kritérií (minimálně rok vydání, jazyk, typ publikace, dostupnost plného textu).
3. Ve výsledkové množině záznamů jsou nalezené duplicitní záznamy z licencovaných EIZ sloučeny do jednoho výsledného záznamu.
4. Discovery systém umožňuje zobrazení online dostupnosti jednotek z knihovního katalogu Zadavatele.
5. Přístup k nalezeným záznamům v jednotlivých EIZ je řešen prostřednictvím OpenURL link serveru, který je součástí dodávky.
6. Discovery systém umožňuje vyhledávání a přístup k výsledkové množině záznamů prostřednictvím zdokumentovaného aplikačního programového rozhraní (API).
7. Uživatelské rozhraní v prostředí WWW je možné graficky přizpůsobit grafickému designu Zadavatele.
8. Uživatelského rozhraní je minimálně v českém a anglickém jazyce.

Licence, která je předmětem veřejné zakázky, bude poskytována studentům, zaměstnancům Zadavatele a čtenářům knihovny Zadavatele (souhrnně uživatelům Zadavatele). Ke dni [den] činí přibližný počet uživatelů Zadavatele [počet] studentů, [počet] FTE akademických pracovníků a [počet] uživatelů z řad veřejnosti, v [název dodavatele] katalogu knihovny Zadavatele je přibližně [počet] záznamů.

6.4 Kritéria a způsob hodnocení nabídek

Základním hodnotícím kritériem je ekonomická výhodnost nabídky.

Dílčí hodnotící kritéria a váhy:

	Dílčí hodnotící kritérium	**Váha v %**
A.	Nabídková cena celkem bez DPH (Licence, technická a provozní podpora)	75 %
B.	Pokrytí zdrojů	15 %
C.	Funkce systému a podpora	10 %

Pro hodnocení nabídek bude použita bodovací stupnice v rozsahu 0 až 100. Každé jednotlivé nabídce je dle dílčího kritéria přidělena bodová hodnota, která odráží úspěšnost předmětné nabídky v rámci dílčího kritéria.

Jednotlivé vypočtené bodové hodnoty dílčích kritérií A až C budou násobeny vahou daného dílčího kritéria. Součet bodových hodnot ze všech dílčích kritérií určí výslednou bodovou hodnotu nabídky. Celkové pořadí nabídek bude sestaveno tak, že nejvýhodnější bude nabídka, která získá nejvyšší celkový počet bodů. V případě rovnosti bodových hodnot dvou či více nabídek rozhoduje o celkovém pořadí nabídek pořadí v kritériu s nejvyšším stupněm významu, tj. kritérium nabídková cena. Veškeré výpočty budou prováděny s přesností na dvě desetinná místa.

6.4.1 Kritérium A: Nabídková cena licence (bez DPH)

Váha kritéria: 75 %

Nabídková cena bez DPH: nabídka s nejnižší nabídkovou cenou obdrží 100 bodů a ostatní nabídky dle vzorce:

$$\frac{\text{nejnižší nabídková cena}}{\text{nabídková cena}} \times 100$$

Dosazuje se číselná hodnota v Kč.

6.4.2 Kritérium B: Pokrytí zdrojů

Váha kritéria: 15 %

Zadavatel bude v rámci tohoto kritéria hodnotit dosaženou míru indexace hodnocených elektronických informačních zdrojů (viz tabulka *Pokrytí indexu* v příloze [číslo přílohy][32]) na základě následujících tří subkritérií:

a) váha zdroje a způsobu pokrytí dat v centrálním indexu
b) míra pokrytí metadaty bez plného textu
c) míra pokrytí metadaty a plným textem

[32] Tabulka *Pokrytí indexu* v příloze Zadávací dokumentace obsahuje seznam poptávaných elektronických informačních zdrojů. Zadavatel předvyplní sloupce Název zdroje, váha$_1$ a váha$_2$. Uchazeč vyplňuje sloupce Smlouva s producentem zdroje (ANO/NE), Míra pokrytí metadaty bez plného textu (0-100 %), Míra pokrytí metadaty s plným textem (0-100 %).

Míru pokrytí jednotlivých hodnocených elektronických informačních zdrojů uchazeč vyplní v Krycím listu pro hodnocení nabídky ve struktuře podle tabulky *Pokrytí indexu*.

1) Ve sloupci „Smlouva s producentem zdroje“ uchazeč vyplní ANO, pokud provozovatel discovery služby má uzavřenu smlouvu přímo s producentem daného zdroje na pravidelné dodávání metadat nebo plných textů pro potřeby centrálního indexu. Pokud jsou data v centrálním indexu vyhledatelná pouze na základě alternativních zdrojů, uvede uchazeč NE.
2) Ve sloupcích míra pokrytí uchazeč uvádí v procentech pokrytí dat daného zdroje v centrálním indexu na úrovni metadat bez plného textu a na úrovni metadat s plným textem.

Součet míry pokrytí v obou sloupcích může být maximálně 100 % (např. při smlouvě s producentem na dodávání všech metadat a plných textů z daného zdroje).

Dílčí míru pokrytí konkrétní periodické publikace v daném zdroji uchazeč vypočte jako podíl počtu let indexovaných v centrálním indexu k počtu let indexovaných daným zdrojem (tj. nepočítá se podle počtu článků, ale podle pokrytí v letech). Dílčí míra pokrytí neperiodické publikace nebo článku bude rovna 0 % při neindexování metadat v centrálním zdroji nebo 100 %, pokud jsou v centrálním indexu zahrnuta minimálně metadata. Pokud publikace není indexována v centrálním indexu, je dílčí míra pokrytí dané publikace 0 %.

Celková míra pokrytí jednotlivého zdroje metadaty/plným textem bude uchazečem vypočtena jako průměr jednotlivých dílčích pokrytí metadaty/plným textem všech periodických a neperiodických publikací daného zdroje, podle toho, jak jsou indexovány v centrálním indexu nabízené discovery služby.

3) Pro publikace s ISBN anebo ISSN, které jsou v centrálním indexu nabízené služby a byly použity pro výpočet míry pokrytí, uchazeč vyplní požadované údaje do tabulky *Seznam*

licencovaných periodických a neperiodických publikací[33], která je v elektronické formě[34] uvedena jako příloha [číslo přílohy] této Zadávací dokumentace. Řádně vyplněnou tabulku s platnými údaji je uchazeč povinen odevzdat ve své nabídce v souboru ve formátu Microsoft Office Excel 2007 (koncovka .xlsx) nebo Microsoft Office Excel 97 - 2003 (koncovka .xls) jako povinnou součást Krycího listu své nabídky. Nedoložení tohoto CD/DVD se souborem podle pokynů této zadávací dokumentace bude mít za následek vyloučení nabídky ze zadávacího řízení.

Údaje v tabulce *Pokrytí zdrojů* a v tabulce *Seznam licencovaných periodických a neperiodických publikací* uchazeč uvede v platném stavu ke dni [den].

Bodové hodnocení:

Ohodnocení pokrytí zdroje, u kterého má provozovatel všechna data indexována centrálním indexem pouze z alternativních zdrojů než je hodnocený zdroj, je vypočítáno podle vzorce:

$$\begin{aligned}&\text{ohodnocení pokrytí jednotlivého zdroje}\\&= \text{váha}_1 \text{ zdroje}\\&\times \frac{\text{Míra pokrytí metadaty bez plného textu } \times\ 0{,}3\ +\ \text{Míra pokrytí metadaty s abstraktem a plným textem}}{100}\end{aligned}$$

Ohodnocení pokrytí zdroje, u kterého má provozovatel minimálně část dat indexováno z centrálního indexu přímo na základě smlouvy s producentem daného zdroje, je vypočítáno podle vzorce:

$$\begin{aligned}&\text{ohodnocení pokrytí jednotlivého zdroje}\\&= \text{váha}_2 \text{ zdroje}\\&\times \frac{\text{Míra pokrytí metadaty bez plného textu } \times\ 0{,}3\ +\ \text{Míra pokrytí metadaty s abstraktem a plným textem}}{100}\end{aligned}$$

[33] Tabulka *Seznam licencovaných periodických a neperiodických publikací* Zadávací dokumentace obsahuje položky k doplnění: ISBN/ISSN, název, producent, nejstarší a nejnovější rok pokrytí metadaty v centrálním indexu, nejstarší a nejnovější rok pokrytí plným textem v centrálním indexu. Tabulka slouží pro kontrolu plnění uzavřené zakázky v případě sporu o uváděné pokrytí metadaty a plným textem.

[34] Použití elektronické formy je vyžadováno vzhledem k předpokládanému velkému rozsahu tabulky vyplněné uchazečem.

Ve výše uvedených vzorcích je míra pokrytí uváděna jako číslo od 0 do 100.

Bodová hodnota hodnocené nabídky je tvořena součtem bodů za všechny hodnocené zdroje:

$$\text{bodová hodnota kritéria} = \sum \text{ohodnocení pokrytí jednotlivého zdroje}$$

Pro dílčí kritérium B. Pokrytí indexu, pro které má nejvhodnější nabídka maximální bodovou hodnotu kritéria, získá hodnocená nabídka takovou výslednou bodovou hodnotu kritéria, která vznikne násobkem 100 a poměru bodové hodnoty kritéria hodnocené nabídky k bodové hodnotě kritéria nejvhodnější nabídky s nejvyšším počtem dosažených bodů.

6.4.3 Kritérium C: Funkce systému a podpora

Váha kritéria: 10 %

Zadavatel bude v rámci tohoto kritéria hodnotit funkce systému a jeho podporu na základě níže uvedených subkritérií podle popisu funkcionality a podpory uvedené uchazečem v nabídce, a to ve struktuře a s bodovými váhami podle tabulky [číslo přílohy] v příloze Zadávací dokumentace[35].

Uchazeč specifikuje hodnocené funkce a podporu popisem v rámci nabídky tak, aby hodnotící komise mohla rozhodnout o míře splnění jednotlivých hodnocených subkritérií. Při stanovení míry splnění daného subkritéria bude hodnotící komisí použito hodnocení *splněno* / *částečně splněno* / *nesplněno*.

Bodová hodnota hodnocené nabídky bude vypočtena jako součet vah a hodnoty míry splnění daných subkritérií, přičemž hodnota míry splnění je dána následujícím způsobem:

splněno: 2 body
částečně splněno: 1 bod
nesplněno: 0 bodů

[35] Tabulka *Funkce systému a podpora* v příloze zadávací dokumentace uvádí výčet hodnocených subkrikérií a stanovené váhy pro hodnocení, viz 6.5 *Závěr kapitoly*. Význam jednotlivých funkcí a způsob jejich hodnocení je vhodné specifikovat samostatně v textu zadávací dokumentace.

Váha jednotlivých subkritérií je uvedena v příloze zadávací dokumentace Funkce systému a podpora.

$$\text{bodová hodnota kritéria} = \sum \text{váha subkritéria} \times \text{hodnota míry splnění subkritéria}$$

6.5 Závěr kapitoly

V případě, že univerzita připravuje výběr centralizované vyhledávací služby indexující i metadata eVŠKP a katalog knihovny, autor doporučuje vycházet při přípravě zadávací dokumentace z doporučení v této kapitole. Uvedená doporučení jsou uvedena tak, aby vysokým školám umožnila minimalizovat náklady na zřízení a provoz discovery služeb a zároveň eliminovala další rizika stanovená v úvodu kapitoly.

Požadovaný způsob doložení míry pokrytí zdrojů zaručuje jasný a transparentní výpočet, zpětně verifikovatelný. Způsob stanovení poptávané funkcionality jednak v části Předmět veřejné zakázky, jednak v Kritériu C hodnocení nabídky, umožňuje oddělit funkcionalitu nutnou (centrální databáze, indexování katalogu a repozitáře eVŠKP, fasetové vyhledávání, deduplikace dat, aplikační rozhraní, vícejazyčné rozhraní včetně podpory češtiny apod.) a funkcionalitu vhodnou (komplexita pokrytí zdrojů, autentizace uživatelů, filtrace záznamů, alerty a RSS, statistiky pro podporu metrik užití, podporu systému apod.).

Subkritéria hodnocení funkcí systému a podpory se uvádějí v příloze zadávací dokumentace. Stanovení váhy jednotlivých subkritérií může knihovna provést např. metodou vážených expertních odhadů.

Doporučená subkritéria pro kritérium Funkce systému a podpora:

1) Autentizace uživatelů (LDAP nebo Shibboleth)
2) Zobrazení aktuální dostupnosti jednotek z katalogu knihovny
3) Možnost filtrace záznamů z knihovního katalogu
4) Možnost filtrace recenzovaných záznamů
5) Nastavení alertů (e-mail, RSS)
6) Export metadat
7) Statistiky využívání

8) Správa prostřednictvím WWW rozhraní
9) Podpora

Díky výběrovému řízení na discovery službu pro Univerzitu J. E. Purkyně v Ústí nad Labem, které vycházelo z autorových doporučení, se podařilo univerzitě získat úsporu 91 % (přes 3,5 mil. Kč bez DPH) oproti původní odhadované maximální částce (4).

7 Vyhledávací rozhraní eVŠKP

Na základě analýzy stávajících rozhraní repozitářů eVŠKP v ČR je zřejmé, že české repozitáře oproti zahraničním zaostávají v možnostech vyhledávání a zpřesňování dotazů. Cílem autora bylo vytvořit modelovou aplikaci vyhledávacího rozhraní eVŠKP, kterou by bylo možné následně využít pro libovolný repozitář eVŠKP s metadaty ve formátu EVSKP-MS. Aplikace má za úkol zajišťovat potřebnou funkcionalitu vyhledávací služby, za účelem implementace pokročilého vyhledávacího rozhraní včetně podpory fulltextového vyhledávání v plných textech českých kvalifikačních prací, v metadatech ve formátu EVSKP-MS a využití faset pro zužování výsledkové množiny záznamů.

Jako experimentální datové základny bylo využito Databáze kvalifikačních prací VŠE. Původní webové rozhraní VŠE v Praze umožňuje pouze jednoduché vyhledávání (viz Obrázek 5 na str. 59) nad metadaty uloženými v relační databázi MySQL. Vzhledem ke způsobu implementace je vyhledávání pomalé (zodpovězení dotazu v řádu až jednotek sekund), neumožňuje zpřesňování dotazu ani fulltextové vyhledávání (fulltextové vyhledávání je podporováno pouze v polích název práce, jméno autora a abstrakt, bez podpory skloňování českých slov).

V rámci výběru vhodné programové platformy provedl autor průzkum dostupných nástrojů umožňujících fulltextové vyhledávání nad bází metadat a plných textů eVŠKP. Vzhledem k požadavku na rychlost, otevřenost řešení a podporu indexace metadatových polí a prací s českým jazykem se jako jednoznačně nejvhodnější ukazuje použití aplikační knihovny Apache Lucene, resp. její nadstavby Apache Solr.

Apache Lucene je aplikační knihovna zajišťující indexaci, ukládání a vyhledávání metadat přes definované aplikační programové rozhraní. Knihovnu Apache Lucene naprogramoval Doug Cutting v roce 1997, od té doby se na vývoji knihovny podílelo téměř sto osob. Knihovna byla zpřístupněna na serveru SourceForge.net, později přešla pod záštitu Apache Software Foundation v rámci projektů Apache Jakarta. V ČR se využitím Apache Lucene pro českou gramatiku (lemmatizací apod.) zabývá několika závěrečných kvalifikačních prací, např. (98) a (99), ze kterých autor disertační práce vycházel při výběru vhodného nástroje.

Pro potřeby modelové aplikace byla zvolena open source podniková platforma pro vyhledávání v metadatech Apache Solr. Aplikace je napsána v jazyce Java, pro indexaci dat a vyhledávání je využito výše zmíněné knihovny Apache Lucene.

V následujících podkapitolách je uveden popis doporučené implementace Apache Solr jakožto vyhledávacího serveru pro plné texty eVŠKP v ČR a metadat ve formátu EVSKP-MS, instalace a konfigurace serveru, redesign uživatelského rozhraní Databáze kvalifikačních prací VŠE.

7.1 Instalace Apache Solr

Apache Solr server byl nainstalován na vývojový server VŠE v Praze ve verzi 4.6.0, která využívá knihovnu Apache Lucene ve verzi 4.6.0. Pro běh aplikačního serveru byl připraven virtualizovaný server s operačním systémem Debian v rámci virtualizačního nástroje Oracle VM VirtualBox. Aplikace Apache Solr je spuštěna v rámci servlet kontejneru Catalina na webovém serveru Apache Tomcat verze 6.

Pro účely testu byla připravena kolekce VSKP, u které byly na základě analýzy metadat eVŠKP autorem disertační práce nakonfigurovány schéma datových polí indexu a ovladač Data Import Handler pro extrakci textů z metadatových záznamů ve standardu EVSKP-MS.

Kolekce VSKP je realizována jako jedna z kolekcí Apache Solr, provozovaného ve variantě multicore (jedna instance aplikace může obsahovat více kolekcí). V popisované implementaci jsou kolekce umístěné do podadresáře cores v rámci hlavního adresáře Apache Solr, soubory kolekce metadat EVSKP-MS jsou umístěny v podadresáři /cores/vskp/.

Jedním ze základních konfiguračních souborů připravených pro potřebu indexace metadat eVŠKP je soubor schema.xml (viz Příloha VII), který obsahuje jednak definici typů polí včetně přiřazených pravidel lemmatizace pro daný jazyk, jednak seznam polí dané kolekce včetně definice názvu pole, typu, způsobu indexování, ukládání a označení, zda pole může nabývat více hodnot zároveň (vhodné pro opakující se prvky, např. volně tvořená klíčová slova uváděná studenty). Kromě polí odpovídajícím prvkům standardu EVSKP-MS jsou v schema.xml definována pole odpovídající různým jazykovým variantám (např. pro dc:title, dcterms:abstract, dc:subject aj. prvky EVSKP-MS).

7.2 Import a extrakce metadat

Import metadat a plných textů do kolekce VSKP je řešen pomocí vestavěné obslužné rutiny, ovladačem „Data Import Request Handler“. Tento ovladač zajišťuje harvestování metadat podle konfigurace v souboru dih-config.xml (viz Příloha IX), který definuje zdroje dat, pravidla extrakce metadat, plného textu a mapování na interní pole indexu. Použití ovladače v kolekci VSKP a umístění souboru s konfigurací ovladače je součástí nastavení kolekce v souboru solrconf.xml (viz Příloha VIII):

```
<requestHandler name="/dataimport"
   class="org.apache.solr.handler.dataimport.DataImportHandler">
   <lst name="defaults">
      <str name="config">dih-config.xml</str>
   </lst>
</requestHandler>
```

Pro import jsou použity dva datové zdroje, v souboru dih-config.xml (viz Příloha IX) označené:

```
<dataSource type="FileDataSource" encoding="UTF-8" />
<dataSource name="dsBinary" type="BinURLDataSource" />
```

První datový zdroj prochází zvolený adresář na diskovém prostoru Apache Solr a načítá z něj jednotlivé XML soubory s metadaty pro import. Zdroj dsBinary slouží pro stažení externího souboru eVŠKP s plným textem ze zadané URL adresy.

V rámci konfigurace extrakce metadat z XML bylo využito komponenty XPathEntityProcessor. Komponenta umožňuje pouze omezené dotazy v syntaxi XPATH, bez podpory jmenných prostorů v atributech, používaných ve formátu EVSKP-MS (konkrétně jednoduchý Dublin Core dc, rozšířený Dublin Core dcterms, pro ETD-MS jmenný prostor thesis a u vlastních prvků jmenný prostor evskp). Toto omezení se projevuje pouze u atributů se jmenným prostorem (např. označení role ve značce <dc:contributor thesis:role="advisor">), kdy není možné tento atribut pro extrakci použít. Chyba v implementaci XPathEntityProcessor byla autorem reportována v rámci systému sledování chyb Apache Solr https://issues.apache.org pod číslem SOLR-5804 (https://issues.apache.org/jira/browse/SOLR-5804).

Řešením uvedeného omezení XPathEntityProcesor je využití XSLT transformace, která odstraní jmenné prostory z importovaného záznamu ve formátu EVSKP-MS.

Použití XSLT transformace uložené v externím souboru dih-remove-ns.xslt (viz Příloha X) je definováno v souboru dih-config.xml (viz Příloha IX) atributem prvku <entity name="xml">:

```
<entity name="xml" processor="XPathEntityProcessor"
   transformer="RegexTransformer,TemplateTransformer,script:checkData"
   datasource="files"
   stream="true"
   forEach="/metadata"
   useSolrAddSchema="false"
   xsl="dih-remove-ns.xslt"
   url="${files.fileAbsolutePath}">
```

V rámci výše uvedené entity jsou definována pravidla extrakce dat ze souboru XML ve formátu EVSKP-MS. Vzhledem k implementaci třídy XPathEntityProcessor (viz výše) a využití XSLT transformace nejsou v pravidlech XPATH uváděny jmenné prostory.

Základní pravidlo použité pro extrakci využívá jazyka XPATH pro označení požadovaného textu pro indexaci, např. pro pole creator označující v kolekci eVŠKP autora práce je uvedena definice:

```
<field column="creator" xpath="/metadata/creator" />
```

V případě extrakce jednoznačného identifikátoru eVŠKP je v případě metadat VŠE v Praze použito transformace regulárním výrazem pomocí třídy RegexTransformer, kdy z pole <identifier> využíváme pouze číselnou část.

```
<field column="id" xpath="/metadata/identifier" regex="(\d+)" />
```

U klíčových slov, která jsou ve standardu EVSKP-MS oddělena čárkou, je definován středník jako dělící znaménko. Rozdělení řetězce je důležité pro následnou tvorbu facet.

```
<field column="subject" xpath="/metadata/subject" splitBy="; " />
```

Formát EVSKP-MS pro zápis dat používá standardu W3CDTF ISO 8601 (http://www.w3.org/TR/NOTE-datetime). V metadatech může být úplné datum ve formátu „*rrrr-mm-dd*“ nebo jen rok vytvoření ve formátu „*rrrr*“. Apache Solr však vyžaduje formát „*rrrr-mm-ddT+00:00Z*“, tj. včetně povinné časové zóny. Pro import dat jsou použity transformace zadané vzorem nebo regulárním výrazem, podle použitého způsobu zápisu dat v XML záznamech Databáze kvalifikačních prací VŠE.

```
<field column="created" xpath="/metadata/created" template="${xml.created}T00:00:00Z" />

<field column="modified" xpath="/metadata/modified"
       regex="(\d{4})-(\d{2})-(\d{2}) (\d{2}):(\d{2}):(\d{2})" />
```

Při testování Apache Solr byla ve verzi 4.6.0 zjištěna chyba v implementaci třídy XPathEntityProcessor, která neumožňuje opakovaný dotaz na stejnou entitu či atribut, chyba byla autorem reportována pod číslem SOLR-5809 (https://issues.apache.org/jira/browse/SOLR-5809).

Ve dvou případech bylo potřeba využít pro transformaci třídy ScriptTransformer, která umožňuje zpracovat data vlastními funkcemi v JavaScriptu, příp. v jiném skriptovacím jazyce podporovaném v jazyce Java 6. Jednotlivé funkce jsou volány pomocnou funkcí function checkData(row). Použití funkcí řeší problém s extrakcí více hodnot z jednoho prvku XML, viz chyba SOLR-5809 popsaná výše.

První funkce, function checkName(row, tag) (viz Příloha IX), zpracovává prvek podle atributu *tag* obsahující dle definice ve standardu EVSKP-MS buď uživatelské jméno ve formátu „*Příjmení, Jméno"* nebo prvek Person dle standardu PersCZ (viz podkapitola 4.1). Příkladem takovýchto XML prvků jsou autor práce a osoba podílející se na zpracování eVŠKP (dc:contributor - vedoucí nebo oponent). Funkce používá mj. pomocných polí se jménem ve formátu attr_*tag*_forename a attr_*tag*_surname, na základě vyhodnocení těchto polí vygeneruje nové pole indexu Apache Solr se jménem ve formátu „*Příjmení, Jméno*".
V případě role přispívající osoby viz výše je použito extrakce hodnoty atributu.

Příklad definice pomocných polí, včetně odstranění tabulátorů a nových řádků regulárním výrazem:

```
<!-- informace o dc:contributor -->
<field column="contributor" xpath="/metadata/contributor"
        regex="([\t\n])" replaceWith="" flatten="false" />
<field column="attr_contributor_role"
        xpath="/metadata/contributor/@role" multiValued="true" />
<field column="attr_contributor_forename"
        xpath="/metadata/contributor/person/name/foreName" multiValued="true" />
```

Druhá funkce function checkUrl(row) (viz Příloha IX) zpracovává prvky evskp:fileProperties a evskp:transfer standardu EVSKP-MS, které jsou ve vzájemné vazbě přes atribut fileID. Funkce vyhledá prvek evskp:fileProperties popisující hlavní práci a k ní odpovídající URL na plný text v prvku evskp:transfer. URL adresa je uložena do pomocného pole attr_url indexu pro následnou extrakci plného textu práce.

Pro extrakci plného textu je v Apache Solr použito komponenty Apache Tika, integrované prostřednictvím třídy TikaEntityProcessor. Studijní informační systém ISIS VŠE požadavky

na plný text na URL adrese s protokolem HTTPS (uvedené v EVSKP-MS XML) přesměrovává na odlišnou URL adresu s plným textem, bohužel implementace TikaEntityProcessor neumožňuje přesměrování při použití protokolu HTTPS. Proto byl autorem disertační práce implementován transparentní proxy server, který plný text z ISIS VŠE na požadavek Apache Solr na pozadí stáhne (včetně potřebného přesměrování) a odešle v odpovědi Apache Solr. Plný text je následně v Apache Solr z PDF extrahován pomocí Apache Tika jako prostý text a zaindexován (alternativně by byla možná i extrakce ve formátu HTML, který však pro naše potřeby fulltextového indexování není vhodný).

V rámci entity je kromě plného textu z metadat extrahován i datový typ a počet stran.

```
<entity processor="TikaEntityProcessor" name="tika" format="text"
        url="http://ciks-test.vse.cz/download/mach/vskp/proxy.ashx?url=${xml.attr_url}"
        dataSource="dsBinary" onError="skip" transformer="RegexTransformer">
  <field name="pdf_contentType" column="Content-Type" meta="true" />
  <field name="pdf_pages" column="xmpTPg:NPages" meta="true" />
  <field name="fulltext" column="text" regex="\n" replaceWith=" "/>
</entity>
```

Kompletní soubor dih-config.xml obsahuje Příloha IX.

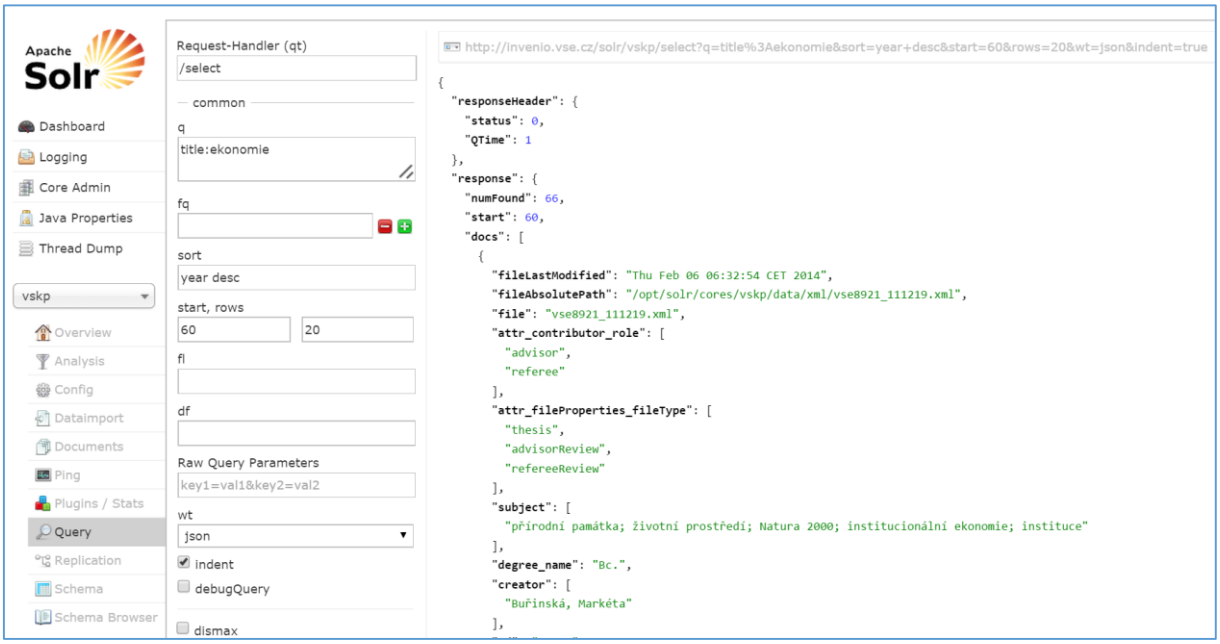
7.3 Uživatelské rozhraní

Apache Solr obsahuje vlastní WWW administrátorské rozhraní, v případě VŠE v Praze dostupné na URL adrese http://solr.vse.cz/solr/[36]. Administrátorské rozhraní je určeno pro konfiguraci serveru, import a testování. Apache Solr podporuje HTTP/REST programové aplikační rozhraní, které je využíváno jak pro dotazování, tak pro správu záznamů a konfiguraci serveru.

Vzhledem k časové náročnosti indexace záznamů probíhalo ladění importu a extrakce polí nejprve na desítkách, později stovkách XML záznamů ve formátu EVSKP-MS. Po dokončení konfigurace serveru bylo zaindexována více jak 30 000 záznamů eVŠKP VŠE v Praze.

[36] Doména http://solr.vse.cz je z bezpečnostních důvodů dostupná pouze v rámci Intranetu VŠE z vyhrazených IP adres (správa Apache Solr, vývojové počítače, WWW server s vyhledávacím rozhraním pro uživatele). TCP port 8983 Apache Solr byl vzhledem k blokování na firewallu VŠE v Praze překonfigurován na povolený TCP port 80 používaný pro HTTP protokol.

Administrátorské rozhraní umožňuje pro testovací účely zkonstruovat dotaz a zobrazit výsledek volání API ve zvoleném formátu (viz Obrázek 19). Odpovídající URL adresa API dotazu je zobrazena uživateli v horní části obrazovky šedou barvou písma.



Obrázek 19 Konfigurace dotazu v administračním rozhraní Apache Solr (zdroj: autor)

Příklad URL dotazu:

http://solr.vse.cz/solr/vskp/select?q=title:work&sort=created+asc&start=150&rows=25&wt=json&facet=true&facet.field=year

Použité parametry v příkladu URL a jejich význam obsahuje Tabulka 4 na následující straně.

Odpověď vyhledávacího serveru se typicky skládá z těchto částí:

a) část responseHeader se stavovým kódem výsledku zpracování dotazu a s dobou potřebnou na zpracování dotazu,
b) část response s počtem nalezených dotazů, pořadovým číslem prvního dokumentu a nalezenými záznamy v opakujícím se vnořeném prvku doc, každý prvek doc obsahuje výčet metadatových polí a konkrétních hodnot (řetězec, pole hodnot apod.),
c) část facet_counts obsahující informace o fasetách a subkritériích faset včetně četností v daném dotazu.

Detailní příklady API dotazů a odpovědí uvádí nápověda Apache Solr (100).

Tabulka 4 Použité parametry v příkladu URL a jejich význam (zdroj: autor)

q=title:work	Vyhledávací dotaz na záznamy obsahující v poli title slovo work (použitý stemmer odpovídající jazyku textu je konfigurován pro každé pole zvlášť).
sort=created+asc	Řazení záznamů podle pole created vzestupně, tj. podle data vytvoření eVŠKP (dcterms:created). Výchozí řazení je podle relevance.
start=150	Získání záznamu ve výsledkové listině s pořadovým číslem 150 a následujících. Parametr start se používá pro stránkování seznamu výsledků v uživatelském rozhraní.
rows=25	Získání 25 záznamů v rámci jedné odpovědi. Pořadové číslo prvního vráceného záznamu je dáno parametrem start.
wt=json	Požadavek na formátování odpovědi ve standardu JSON. Alternativně lze získat odpověď ve formátech XML, PHP, Ruby, CSV aj.
facet=true	Zapnutí funkce facet.
facet.field=year	Seznam požadovaných facet, v tomto případě rok obhajoby z metadatového pole EVSKP-MS dcterms:dateSubmitted.

Uživatelské rozhraní je řešeno jako samostatná webová aplikace, která využívá Apache Solr jako fulltextový vyhledávací server. Vzhledem k záměru provést redesign uživatelského rozhraní Databázi kvalifikačních prací VŠE, autor disertační práce zvolil programovací jazyk PHP a šablonovací systém Smarty, shodně s vývojovým prostředím Databáze kvalifikačních prací VŠE.

Uživatelské vyhledávací rozhraní v PHP je dostupné na webovém serveru s povoleným HTTP(S) přístupem k Apache Solr. Webový server zpracuje požadavek uživatele na vyhledávání s výběrem kritérií faset pro omezení dotazu. V PHP je volána URL adresa podle specifikace Apache Solr API a z vyhledávacího serveru stažena odpověď ve formátu JSON, na základě které je uživateli vygenerována HTML stránka s interpretovanými výsledky.

Při návrhu webového rozhraní autor volil mezi různými koncepty zobrazování faset a jejich kritérií, přičemž vycházel z charakteristik repozitářů VŠKP podle průzkumu repozitářů v podkapitole 2.3. Ve všech třech příkladech může uživatel zadat vyhledávané slovo, které je vyhledáváno ve všech indexovaných metadatových polích, a k tomu dotaz zúžit výběrem kritérií pro zobrazené fasety.

Koncept první odpovídá Národnímu úložišti šedé literatury. Pokud uživatel vybere omezení výsledkové množiny záznamů volbou kritéria z fasety (např. klíčová slova: likvidita), kritéria fasety v následujícím kroku odpovídají zúžené výsledkové množině záznamů. Kritéria použitá v rešerním dotazu jsou zobrazena v záhlaví levého menu s fasetami, kde je možné je selektivně zrušit. Pokud výsledková množina záznamů již neobsahuje jiná kritéria z dané fasety, faseta se v menu nezobrazuje. V jednotlivých krocích lze zpřesňovat např. požadovaná klíčová slova v jednom záznamu a tím zužovat rešeršní dotaz (např. klíčové slovo: likvidita AND rentabilita), ale nelze dotaz rozšířit (např. klíčové slovo: likvidita OR rentabilita). Shrnutí: Více kritérií ze stejné fasety je v rešerním dotazu spojeno booleovským operátorem AND, více faset booleovským operátorem AND. V jednom kroku rešerního dotazu lze vybrat nebo odebrat pouze jedno kritérium jedné fasety.

Koncept druhý odpovídá přístupu Repozitáře závěrečných prací Univerzity Karlovy. Uživatel může v jednom kroku vybrat více kritérií (např. rok obhajoby: 2015 OR 2014). Po upřesnění výběru a načtení nové stránky jsou v rámci dané fasety zobrazeny jen dříve vybrané kategorie (tj. jen kategorie odpovídající výsledkové množině záznamů), zaškrtnuté, uživatel může pouze kritéria ve fasetě zrušit a tím dotaz zúžit (např. odebrat z dotazu rok 2014, výsledné omezení je rok obhajoby: 2015). Rešeršní dotaz nejde v jednom kroku pro danou fasetu rozšířit (např. rok obhajoby: 2015 OR 2014 OR 2013). Kategorie faset nezobrazují počet souvisejících záznamů, uživatel tedy nemůže predikovat výsledek své volby. Repozitář neobsahuje fasety, kde by bylo možné u dokumentu mít více variant, např. jména autorů nebo klíčová slova. Shrnutí: Více kritérií ze stejné fasety je v rešerním dotazu spojeno booleovským operátorem OR, více faset booleovským operátorem AND. V jednom kroku rešerního dotazu lze vybrat nebo odebrat více faset a kritérií.

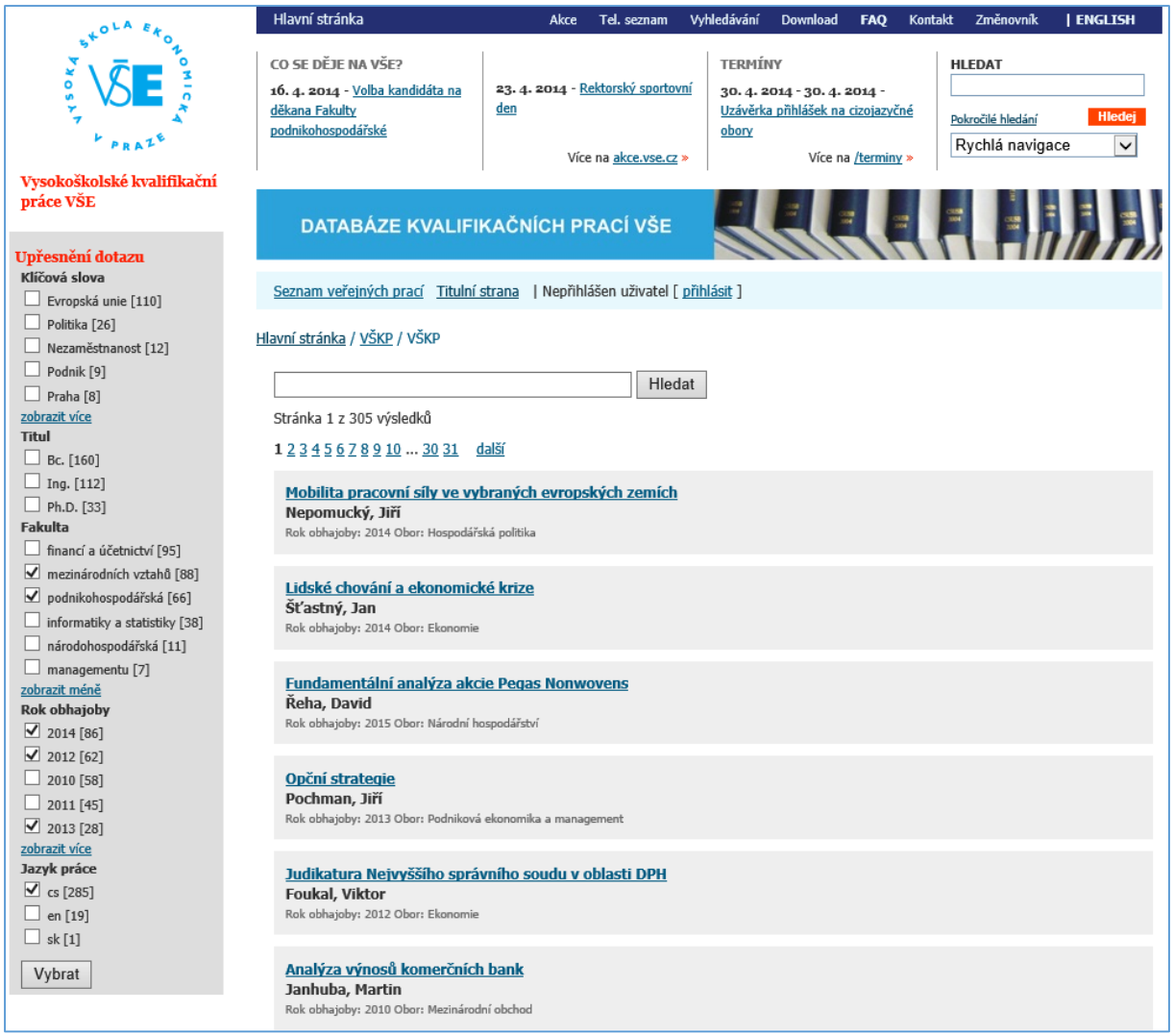
Třetí koncept odpovídá přístupu projektu DART-Europe E-theses Portal. Po vybrání kritéria a načtení nové výsledkové množiny záznamů zůstává faseta zobrazena včetně zaškrtnutého zvoleného kritéria, které je možné následně opět z rešerního dotazu odebrat. Kromě

zvoleného kritéria jsou ale v rámci fasety zobrazena i kritéria alternativní (nenacházející se ve výsledkové množině záznamů zúžené fasetami), takže uživatel může rešeršní dotaz rozšířit přidáním dalšího kritéria v dané fasetě (např. rok: 2015 OR 2014). V závorce u kritérií je uveden počet záznamů, které zahrnutím kritéria do rešeršního dotazu budou přidány do výsledkové množiny záznamů.

Z uvedených konceptů byly připraveny prototypy vyhledávacích rozhraní a provedeno uživatelského testování, na základě kterého byla pro Databázi kvalifikačních prací VŠE zvolena obdoba varianty třetí. Více kritérií ze stejné fasety je v rešerním dotazu spojeno booleovským operátorem OR, více faset booleovským operátorem AND. V jednom kroku rešerního dotazu lze vybrat nebo odebrat více kritérií jedné fasety.

Kritéria uvedená v rámci faset odpovídají vyhledávacími dotazu uživatele nad všemi prohledávatelnými metadaty včetně plného textu. Po vybrání kategorií faset a opětovném načtení stránky s aktualizovanou výsledkovou množinou záznamů jsou uživateli zobrazeny fasety a kritéria odpovídající výchozímu dotazu bez aplikování faset. Kritéria faset zahrnutá do rešerního dotazu jsou zaškrtnuta, uživatel může pro zúžení či rozšíření dotazu kritéria libovolně přidávat nebo odebírat. Více kritérií zvolených ze stejné fasety je v rešerním dotazu spojeno booleovským operátorem OR, více faset booleovským operátorem AND. V jednom kroku rešerního dotazu lze vybrat nebo odebrat více faset a kritérií.

V případě velkého počtu kritérií v rámci jedné fasety jsou zobrazeny ve výchozím stavu jen kritéria s nejvyšší četností záznamů, ostatní kritéria může uživatel zobrazit na vyžádání (autor naprogramoval odpovídající funkci v JavaScriptu pro zobrazení/skrytí kritérií bez nutnosti znovunačtení stránky). Navržené uživatelské rozhraní s využitém faset pro Databázi kvalifikačních prací VŠE zobrazuje Obrázek 20.



Obrázek 20 Uživatelské rozhraní s fasetovou navigací (zdroj: autor)

Závěr kapitoly

Autor pro potřeby vyhledávání nad metadaty eVŠKP vybral a doporučuje použití platformy Apache Solr, která spojuje výhody knihovny Apache Lucene a nadstavby Apache Solr s HTTP/REST aplikačním programovým rozhraním.

V této kapitole autor popsal konkrétní způsob instalace a konfigurace Apache Solr pro potřeby indexování a vyhledávání metadat a plných textů eVŠKP. V souboru dih-config.xml (viz Příloha IX) je řešena konkrétní extrakce polí z metadat ve formátu EVSKP-MS. Vzhledem k chybám, které byly nalezeny v testované verzi Apache Solr, autor navrhuje dočasná řešení problémů, která účinně umožňují extrakci dat i přes daná omezení.

Na příkladu metadat z Databáze kvalifikačních prací VŠE byl otestován import více jak 30 000 metadatových záznamů a plných textů eVŠKP. V testech vyhledávání provedených autorem se potvrdila velmi rychlá odezva indexovacího serveru, kdy výsledková množina záznamů byl vygenerována, i v případě komplikovanějších dotazů, za přibližně 150 ms. Při opakování stejného dotazu, díky využití interní cache Apache Solr, je výsledek vygenerován do 10 ms. Pro porovnání, vyhledání požadovaných záznamů v databázovém serveru MySQL Databáze kvalifikačních prací VŠE (používané v původním výhledávacím rozhraní) trvá v řádu stovek milisekund, podle druhu dotazu.

Redesignované uživatelské rozhraní na bázi PHP umožňuje využití faset (v případové studii použity fasety Klíčová slova (pro metadata volně tvořená uživateli), udělovaný akademický titul, Fakulta, Rok obhajoby, Jazyk práce) a jejich postupné přidávání a odebírání za účelem zpřesnění či zúžení rešeršního dotazu. Číselná hodnota uvedená u kategorií faset udává počet záznamů odpovídajících dané kategorii.

Vzhledem k široké nabídce funkcí umožňuje Apache Solr tvorbu různých vyhledávacích rozhraní včetně velmi komplexní podpory faset. V případě, že by už tak velmi rychlé vyhledávání nepostačovalo potřebám organizace, Apache Solr nabízí široké možnosti škálování a distribuovaného vyhledávání.

Na základě výše uvedených zkušeností autor pro praxi v ČR doporučuje Apache Solr jako ověřené řešení pro tvorbu vyhledávacích rozhraní nad bázemi metadat a plných textů, ať již se jedná o české repozitáře eVŠKP, nebo o Open Access repozitáře odborných vědeckých publikací (kde lze využít i některé z nabízených aplikací, postavených na bázi knihovny Apache Lucene anebo platformy Apache Solr).

8 Plagiátorství u eVŠKP

Repozitáře eVŠKP ČR jsou pro publikující autory, včetně studentů, nejen velmi dobrým informačním pramenem pro podkladové materiály, ale také snadným zdrojem pro kopírování plných textů. Nejčastěji jsou eVŠKP zneužívány studenty při psaní seminárních a závěrečných kvalifikačních prací na příbuzné téma.

Vedoucí VŠKP a oponenti mají důležitou roli v kontrole originality předkládaných prací k obhajobě, zda se student nepokouší vydávat plagiát za své dílo. Aby nemuseli jednotlivé práce ručně procházet ve vyhledávači a tím ověřovat, zda předkládaný text je originální myšlenkou autora, parafrází jiného zdroje či dokonce přímou citací cizího textu bez korektního uvedení zdroje, je žádoucí nad repozitářem studentských prací implementovat softwarovou podporu vyhledávání duplicit v textech. Analýzou vhodných nástrojů se zabývá tato kapitola[37].

Nástroj na podporu vyhledávání duplicit v textech musí provádět kontrolu napříč různými zdroji, ze kterých studenti čerpají informace při zpracování svých seminárních prací a VŠKP. Na základě zpracovaných analýz plagiátů eVŠKP v uplynulých letech autor disertační práce může konstatovat, že studenti při plagiátorství využívají především volně přístupné webové stránky na Internetu, již obhájené a na Internetu v plném textu dostupné eVŠKP a dříve zpracované vlastní seminární práce[38]. V menší míře využívají odborné texty z elektronických informačních zdrojů dostupných na univerzitě či přepisují texty např. z učebnic a skript.

Pro potřeby univerzit v ČR je k dispozici několik systémů na kontrolu duplicit v repozitářích eVŠKP. Jedním z nich je i český systém Theses.cz vyvinutý na Masarykově univerzitě (viz oddíl 2.3.2). Ze zahraničních systémů jsou na vysokých školách nejčastěji používány aplikace

[37] Text popisující zpracovanou analýzu vychází z prezentace a příspěvku autora publikovaného v roce 2013 ve sborníku 6. ročníku Semináře ke zpřístupňování šedé literatury (119). Na podzim 2014 byla zveřejněna nová verze Theses.cz, která zásadněji upravuje zobrazování podobností menších než 5 % textu a nově umožňuje výpočet celkové podobnosti. Byly tak odstraněny dvě zásadní vlastnosti Theses.cz negativně hodnocené v tomto testu.

[38] V případě korektní autocitace vlastní, dříve odevzdávané práce se nejedná o plagiát, rozšíření seminární práce na závěrečnou kvalifikační práci může být vyučujícím schváleno. Studenti však často přebírají svůj dříve obhajovaný text bez vědomí vyučujícího a neuvádějí citace svých původních vlastních prací.

Ephorus a Turnitin. Pro správce a vedení univerzit je tak důležitá otázka, jaký systém je vhodný pro repozitáře vysokoškolských kvalifikačních prací v ČR.

V této kapitole se autor disertační práce zabývá komparativní analýzou a testem systémů na odhalování duplicit (tzv. antiplagiátorských systémů) a na prototypovém návrhu vlastní aplikace ukazuje potenciál vyhledávání duplicit eVŠKP vůči volně dostupným zdrojům na Internetu vyhledaným pomocí vyhledávače Google.

Pro potřeby testu byl v úvodu vytvořen textový korpus obsahující nejčastější zdroje, ze kterých studenti čerpají, a byly nasimulovány úpravy prováděné plagiátory. Zároveň byly stanoveny hypotézy chování ideálního nástroje na vyhledávání duplicit.

Následně byla na připraveném korpusu experimentálně ověřena úspěšnost detekce duplicit u nejvýznamnějších antiplagiátorských systémů, provedeno jejich komparativní srovnání a ověření stanovených hypotéz. Součástí evaluace byla také vlastní prototypová aplikace autora na odhalování plagiátů využívající vyhledávače Google. Na základě výsledků jsou navrženy vhodné úpravy aplikací.

8.1 Tvorba textového korpusu

Přes celkem velké množství nástrojů pro detekci duplicit v dokumentech neexistuje jednoznačná metodika na jejich hodnocení. Jedním z hlavních důvodů je nejednoznačnost, jak by měl vypadat textový korpus pro evaluaci těchto nástrojů – z jakých textů by měly vzniknout ukázkové dokumenty, na kterých bude prováděno vyhledávání duplicit. V případě vysokoškolských kvalifikačních prací je navíc zdroj plagiátorství prakticky neomezený, plagiátor může využít dalších studentských prací, zdrojů na Internetu volných či placených i monografií.

Pokud bychom dokázali identifikovat vhodné zdroje pro textový korpus a testovací data, pro potřeby veřejného testovacího korpusu je potřeba svolení držitelů autorských práv takovýchto textů. V případě interního užití pro účely vědeckého výzkumu bez zveřejnění plných textů, jako bylo u této studie, můžeme vycházet ze zákonné licence Autorského zákona (7), konkrétně využití zákonné licence pro citace v § 31.

Pro účely studie zabývající se plagiátorstvím u vysokoškolských kvalifikačních prací nebylo možné použít žádný z běžně dostupných korpusů, neboť jsou určeny většinou pro anglické texty[39] nebo pokrývají pouze specializované zdroje[40].

Z výše uvedených důvodů byl pro test připraven vlastní textový korpus obsahující celkem 300 vzorků textu založený na 50 dokumentech (přesnější výsledky by poskytl výrazně větší korpus z více zdrojů, v řádu stovek dokumentů celkem). Pro podporu přípravy korpusu byla naprogramována vlastní aplikace (viz Obrázek 21) pro evidenci zdrojových dokumentů (texty a metadata ve formátu XML), automatizované transformace textů, generování testovacích dat v HTML a generování statistik pro Excel.

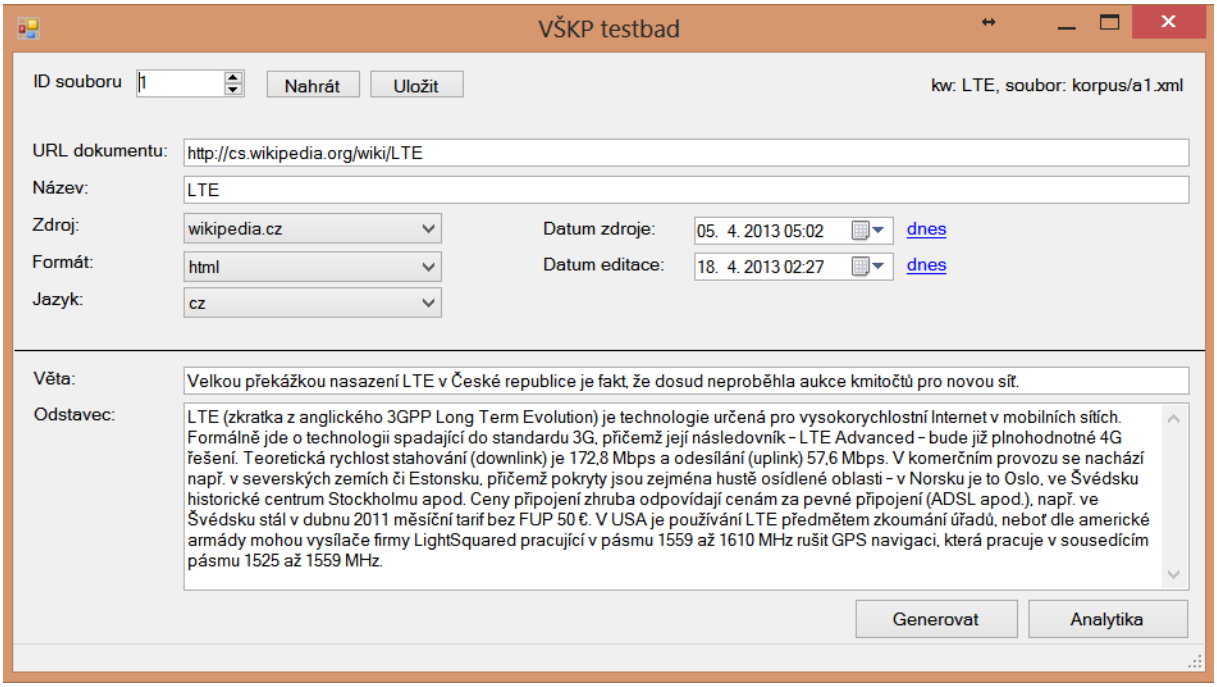
Následující oddíly této podkapitoly popisují použité zdroje, metodiku výběru dokumentů a provedené transformace textů, které simulují různé způsoby plagiátorství zmíněné v definici pojmu plagiátorství v oddílu 2.1.5 disertační práce.

Pro zajištění objektivity provedených testů bylo potřeba zvolit náhodná data ze zdrojů, které studenti používají při psaní vysokoškolských kvalifikačních prací. Pro testovací vzorek bylo stanoveno deset zdrojů a z každého vybráno pět dokumentů na předem daná klíčová slova.

Jako klíčová slova byly vybrány zkratky, které se používají v českém i anglickém jazyce. Díky tomu je výběr dokumentů na jazyku neutrální. Při výběru dokumentů do textového korpusu byl z každého zdroje preferován první nalezený záznam s plným textem, neboť i studenti při plagiátorství většinou volí mezi prvními z dostupných dokumentů. Pokud systém neumožňuje řazení nalezených výsledků podle relevance, je použito řazení od nejnovějších po nejstarší dokumenty.

[39] Příkladem českých textů byly např. korpusy psaného a mluveného jazyka v Českém národním korpusu (ČNK). Až po uzavření tvorby testovacího korpusu této studie byl v roce 2013 jako součást ČNK zveřejněn korpus SKRIPT2012 obsahující texty školních písemných prací, který jako jediný by mohl být použit jako vhodný dílčí zdroj dat pro testy.

[40] Příkladem specializovaného zdroje je datový korpus ClueWeb99 používaný pro soutěž systémů na vyhledávání plagiátorství v rámci konference PAN 2013. Účastníci soutěže používají k vyhledání zdrojových dokumentů předem definované webové aplikační programové rozhraní k tomuto korpusu, cílem úkolu je minimalizovat náklady na stažení (minimalizovat počet stažených souborů).



Obrázek 21 VŠKP testbad - aplikace pro přípravu testovacího korpusu (zdroj: autor)

Klíčová slova

Pro zajištění náhodnosti testu autor záměrně vybral téma mobilních komunikací, které nemá spojitost s knihovní vědou, kde lze u studentů očekávat vyšší informační gramotnost. Volba tématu a klíčových slov ovlivňuje výběr dokumentů z jednotlivých zdrojů, nepředpokládáme však zásadní vliv klíčových slov na následnou úspěšnost detekce projevů plagiátorství.

Autor stanovil následujících 5 termínů – zkratek, které se používají v pracích na zvolené téma jednotně jak v českém, tak anglickém jazyce:

1. LTE
2. UMTS
3. EDGE
4. WiMax
5. Wi-Fi

Při výběru textů se u klíčového slovo EDGE ukázalo, že vyhledávače většinou nacházely články obsahující anglické slovo edge (v českém významu: hrana, okraj) než zkratku EDGE psanou velkými písmeny (angl. Enhanced Data rates for GSM Evolution). To pro relevantnost výběru nevadí, a proto v testovacím korpusu jsou např. i články o seriálu Edge of Darkness nebo recenze filmu Bridget Jones: The Edge of Reason.

8.1.2 Zdroje dat a použité vyhledávače

Výše uvedených pět termínů bylo vyhledáno v deseti českých a anglických zdrojích, ze kterých studenti většinou kopírují:

A. česká Wikipedie
B. anglická Wikipedie
C. vysokoškolské kvalifikační práce ČR psané česky
D. vysokoškolské kvalifikační práce ČR psané anglicky
E. mezinárodní vysokoškolské kvalifikační práce
F. monitoring českých médií
G. mezinárodní odborné práce v režimu Open Access
H. české webové stránky
I. anglické webové stránky
J. elektronické informační zdroje

V každém zdroji byly vyhledány dokumenty na výše uvedená klíčová slova. V textu kapitoly je odkazováno na tyto dokumenty pomocí kódu v hranatých závorkách, složeného z písmene označujícího zdroj a čísla označujícího klíčové slovo. Seznam všech použitých dokumentů v testovacím korpusu a odpovídajících kódů odkazů je uveden v Příloze XII.

Wikipedie

Wikipedie patří u studentů mezi jeden z nejčastěji používaných zdrojů při psaní, a to i přes to, že se jedná povětšinou o zdroj sekundární, nedůvěryhodný a tím i nevhodný pro VŠKP. Wikipedie je také preferována vyhledávačem Google, kde odkaz na Wikipedii je většinou na prvním místě výsledkové množiny záznamů – o to více je využívána méně informačně gramotnými studenty dopouštějícími se plagiátorství. Do testovacích dat byly zahrnuty části z českých a anglických stránek Wikipedie, celkem 20 % z celého testovacího souboru.

Vysokoškolské kvalifikační práce ČR

Dalším oblíbeným zdrojem při psaní VŠKP, který je vhodné do testu zahrnout, jsou dříve obhájené práce na shodné téma. Díky národnímu registru vysokoškolských kvalifikačních prací Theses.cz je mnoho prací snadno dohledatelných a tak i snáze dostupných pro kopírování, ale je také pro zúčastněné školy zajištěna kontrola obhajovaných eVŠKP vůči již zaindexovaným pracím. Plné texty z Theses.cz nejsou automatizovaně dostupné pro indexaci dalšími vyhledávacími nástroji ani většinou nejsou volně dostupné z webů univerzit, je proto

zhoršena možnost detekce duplicit z českých eVŠKP v antiplagiátorských nástrojích, které nebyly vyvinuty Masarykovou univerzitu.

Při výběru dat do testovacího korpusu byla z Theses.cz vybrána první nalezená česká anebo anglická kvalifikační práce pro dané klíčové slovo, kde byl plný text volně přístupný na Internetu. V celém testovacím korpusu je 20 % českých kvalifikačních prací, z toho polovina českých a polovina anglických. Z deseti prací byla jedna obhajována v roce 2013, tři v roce 2012, ostatní práce jsou staršího data.

Z výše uvedených důvodů můžeme stanovit hypotézu, že právě systém Theses.cz by měl v detekci duplicit u českých eVŠKP dosahovat nejlepších výsledků.

Mezinárodní vysokoškolské kvalifikační práce

Zahraniční kvalifikační práce studenti většinou pro bakalářské a diplomové práce nevyužívají kvůli neochotě ke studiu zdrojů v cizím jazyce. Využití cizojazyčné eVŠKP je pravděpodobnější u prací disertačních, kde již lze předpokládat vyšší informační gramotnost autorů a tím i nižší riziko plagiátorství.

Nejvýznamnějším zdrojem volně dostupných vysokoškolských kvalifikačních prací na mezinárodní úrovni je systém Networked Digital Library of Theses and Dissertations (NDLTD), resp. aplikace SCIRUS ETD Search (101) pro vyhledávání v univerzitních repozitářích zapojených do NDLTD. Preferovány byly první nalezené výsledky na dané klíčové slovo. Samotné práce jsou kromě repozitářů škol většinou publikovány, často v ProQuest Dissertations & Theses Database.

V testovacím korpusu je zahrnuto 5 záznamů nalezených systémem SCIRUS ETD Search, z toho jedna práce obhájená v roce 2012 a čtyři staršího data.

Monitoring českých médií

Jedním z méně využívaných zdrojů, i když velmi významným, jsou média. Pro vyhledávání tohoto typu dokumentů bylo použita databáze Anopress, vybírány byly články vydané v posledních 12 měsících, na dané klíčové slovo, první nalezený. Jeden z článků je slovensky, ostatní čtyři česky.

Vzhledem k tomu, že monitorované zdroje nejsou vždy dostupné online, existuje hypotéza, že u tohoto zdroje bude nízké procento z celkového počtu nalezených podobností oproti zdrojům volně dostupným na Internetu.

Mezinárodní odborné práce v režimu Open Access, elektronické informační zdroje

U informačně gramotnějších uživatelů, tj. především u disertačních prací a v odborných vědeckých článcích lze očekávat větší míru využití (volně dostupných) odborných vědeckých článků. Do testu byly zahrnuty zahraniční recenzované práce volně dostupné na Internetu (10 % testovacího korpusu), tj. publikované v režimu Open Access, a práce publikované v odborných recenzovaných časopisech (dalších 10 % testovacího korpusu).

Pro vyhledání volně dostupných prací byl použit systém arXiv.org. Obsahuje přes 800 tisíc Open Access článků mj. z fyziky a počítačové vědy, tj. oborů, kde se nacházejí i články na stanovené téma mobilních komunikací. Při výběru článků z arXiv.org bylo vyhledáváno napříč všemi obory, vybrán byl první nalezený záznam. Systém primárně řadí od nejnovějších článků po nejstarší, proto v testovací sadě z pěti záznamů jsou čtyři z roku 2013 a jeden z konce roku 2012.

Při vyhledávání tištěných odborných dokumentů byly použity elektronické informační zdroje dostupné na Vysoké škole ekonomické v Praze, v každém zdroji vybrán článek na jedno z klíčových slov podle následujícího rozdělení klíčové slov / zdroj:

1. LTE: EBSCO
2. UMTS: ProQuest Central
3. EDGE: eBrary (kolekce eBooků na téma Business & Economics)
4. WiMax: JSTOR
5. Wi-Fi: OECD iLibrary (plné texty publikací vydaných OECD)

V případě EIZ lze předpokládat, že texty zde uvedené mají vyšší odbornou úroveň, nevýhodou pro kontrolu duplicit je uzavřenost daných zdrojů pro internetové vyhledávače. Odhalení plagiátorství bude možné spíše díky sekundárním zdrojům, pre/postprintům zveřejněným autorem nebo článkům v režimu Gold Open Access. Oproti tomu Open Access zdroje indexované v arxiv.org mají výhodu dostupnosti plného textu vyhledávačům, nevýhodou je umístění na spodních místech v seznamu nalezených výsledků (důvodem např. nižší PageRank lokálních repozitářů u Google). V analýze bude zajímavé sledovat, jak

si s těmito dvěma druhy zdrojů (Open Access vs. EIZ) poradí jednotlivé vyhledávací nástroje. Lze předpokládat, že lepších výsledků obecně u těchto zdrojů dosáhnou zahraniční antiplagiátorské nástroje oproti českým.

Webové stránky

Posledním zástupcem zdrojů používaných při plagiátorství jsou webové stránky. Přes to, že se jedná většinou o méně důvěryhodné zdroje, webové stránky jsou nejčastějším zdrojem textů pro plagiátory. Schopnost vyhledávat ve webových stránkách je proto pro dobrý antiplagiátorský systém zásadní. Předpokládáme, že velmi dobrých výsledků budou dosahovat systémy využívající služeb webových vyhledávačů.

V českém prostředí jsou dominantní dvě velké vyhledávací služby, Seznam a Google. Volba vyhledávače je pro kvalitu detekce duplicit důležitá, neboť autoři vysokoškolských kvalifikačních prací čerpají z dokumentů nalezených jako první a míra indexace webu i řazení dokumentů ve výsledkové listině hledání se u uvedených služeb zásadně liší.

Podle průzkumu společnosti Effectix (102) čeští uživatelé volí častěji Google, proto byl tento vyhledávač v česká a anglické verzi použit pro výběr webových dokumentů na daná klíčová slova. Preferován byl první nalezený dokument s delším plným textem. V celém korpusu je 10 % webových stránek nalezených v českém jazyce a 10 % v jazyce anglickém, tj. jedna pětina všech dokumentů.

8.1.3 Použité transformace

Z dokumentů vyhledaných podle výše uvedených pravidel byla náhodně vybrána jedna věta ze začátku článku, obsahující pokud možno souvislý text s daným klíčovým slovem bez závorek a bez horních indexů u poznámek pod čarou apod. Dále byl z dokumentu náhodně vybrán odstavec textu o velikosti přibližně 5 vět. Případné odkazy na poznámky, zdroje apod. byly odstraněny. Uvedené fragmenty textu byly společně s popisnými metadaty o dokumentu uloženy do databáze výše zmíněné aplikace vyvinuté pro přípravu testovacích dat.

Úkolem testu bylo ověřit následující hypotézy o systémech na odhalování plagiátů:

1. Aplikace umí odhalit jednu větu zkopírovanou ze zdrojového dokumentu.
2. Aplikace umí odhalit jeden odstavec zkopírovaný ze zdrojového dokumentu. Aplikaci nevadí případná zalomení řádků, indexy apod. ve zdrojovém nebo testovaném dokumentu.
3. Pro úspěšnou detekci nevadí, pokud plagiátor přidá/odebere slovo v kopírované větě.
4. Aplikace provádí detekci českých textů nezávisle na diakritice.
5. Pro úspěšnou detekci nevadí, pokud plagiátor parafrázuje jedno slovo ve větě.
6. Pro úspěšnou detekci nevadí, pokud plagiátor parafrázuje celou větu.
7. Pro úspěšnou detekci nevadí, pokud plagiátor přeloží text z/do českého jazyka.

V rámci aplikace byla provedena řada transformací vložených vět, které měly za úkol simulovat výše uvedené metody plagiátorství. Pro automatický překlad a parafrázi slov/vět bylo použito HTTP/REST programové rozhraní aplikace Microsoft Translator (103), konkrétně metody Translate a Paraphrase API. Rozhraní Microsoft Translator API, s omezeným počtem dotazů denně zdarma, bylo upřednostněno před obdobnou funkcionalitou Google Translate API, která je zpoplatněna. Prováděné transformace zdrojové věty se lišily pro české texty a texty v cizím jazyce.

Testovací korpus dat generovaný aplikací obsahoval celkem 300 fragmentů textu získaných z výše uvedených zdrojů a za využití níže uvedených transformací.

Texty v českém jazyce (19 dokumentů, 38 % dat)

- jedna věta ze zdrojového dokumentu, bez úprav,
- jeden odstavec ze zdrojového dokumentu, bez úprav,
- věta se dvěma slovy prohozenými,
- věta s odstraněnou diakritikou,
- věta s jedním slovem nahrazeným významově blízkým (parafráze slova),
- věta přeložená automaticky do anglického jazyka.

Texty v cizím jazyce (31 dokumentů, 62 % dat)

- jedna věta ze zdrojového dokumentu, bez úprav,
- jeden odstavec ze zdrojového dokumentu, bez úprav,
- věta se dvěma slovy prohozenými,
- věta přeložená automaticky do českého jazyka,
- věta s jedním slovem nahrazeným slovem významově blízkým (parafráze slova),
- věta s více slovy nahrazenými slovy významově blízkými (parafráze věty).

Aplikací pro správu textového korpusu byly vygenerovány HTML soubory s texty, které byly buď přímo, nebo po konverzi do formátu Word, nahrány do jednotlivých testovaných antiplagiátorských systémů.

8.2 Zhodnocení antiplagiátorských systémů

V rámci testu byla porovnána funkcionalita nejvýznamnějších systémů používaných pro kontrolu plagiátorství v závěrečných studentských pracích v České Republice – aplikace Turnitin, Ephorus a systémy Masarykovy univerzity (Theses.cz, Odevzdej.cz).

V rámci komparační analýzy byly výsledky výše uvedených systémů porovnány s výsledky aplikace GooglePlagiarism vyvinuté autorem této práce.

8.2.1 Turnitin

Základní údaje

Antiplagiátorský systém Turnitin http://www.turnitin.com je vlastněn společností iParadigms, LLC, USA. Lokální zastoupení pro ČR poskytuje vedení společnosti iParadigms Europe Ltd., Velká Británie, kontakt pro ČR: Markéta Vágnerová, e-mail: mvagnerova@turnitin.com. Testování programu probíhalo v rámci měsíčního trialu Vysoké školy ekonomické v Praze ve dnech 15. 5. – 14. 6. 2013. V rámci testování autor disertační práce zorganizoval dne 10. 6. 2013 na VŠE v Praze setkání se zástupci společnosti s prezentací aplikace.

Systém Turnitin byl vyvinut na základě recenzního softwaru a prototypu aplikace Turnitin na detekci neoriginálních prací studentů. Podle webových stránek projektu (104) bylo v roce 2012 ve společnosti zaměstnáno 120 lidí, systém Turnitin zpracoval přes 80 milionů dokumentů s průměrnou dobou zpracování dokumentu 13 sekund (při testu aplikace reálná doba od nahrání do zpřístupnění výsledků kontroly byla okolo 30 sekund). Systém je k dispozici v 15 jazykových mutacích bez češtiny. Systém využívá přes 1 milion aktivních instruktorů a 20 milionů studentů z 10 000 vzdělávacích institucí.

Podle informací společnosti je pro porovnávání k dispozici databáze s více jak 24 miliardami webových stránek, 300 miliony archivovaných studentských prací a 120 miliony článků z více jak 110 000 časopisů a knih.

Licence je poskytována na bázi počtu studentů zapsaných do denního prezenčního studia na dané instituci. Cena roční licence pro univerzity je bez slev £1 430 plus příplatek £1,16 za každého studenta, který instituci navštěvuje (v případě integrace s Moodle či jiným z nabízených online vzdělávacím nástrojů je příplatek £0,23 za studenta). Výsledná cena aplikace Turnitin pro VŠE v Praze je po započtení slev okolo GBP 19 000, tj. přes půl milionu Kč.

Aplikace Turnitin umožňuje spravovat dokumenty přes webové rozhraní nebo napojením na některý z externích systémů (např. Moodle). Rozlišuje tři uživatelské skupiny – správce instituce, vyučující a studenty. Jedna osoba může nabývat jedné, dvou nebo všech tří rolí (např. doktorand, který vede kurzy a je zároveň správcem za katedru). Správci především přidělují oprávnění přístupu dalším uživatelům. Vyučující mohou zakládat jednotlivé kurzy a související odevzdávárny. Do připravených odevzdáváren studenti s přístupovým heslem nahrávají soubory, nebo tak může učinit za studenty sám vyučující.

Kromě kontroly plagiátorství obsahuje Turnitin moduly GradeMark (hodnocení prací vyučujícími, s možností připojení poznámek k textu se zpětnou vazbou studentovi) a PeerMark (hodnocení prací studenty navzájem s možností dotazníku recenzentům a anonymních recenzí, podporující kritické myšlení u studentů).

Způsob odevzdání

Při zakládání odevzdávárny vyučující stanovuje následující údaje pro odevzdávání a vyhledávání duplicit:

- přístupový kód pro odevzdávání,
- poslední den pro odevzdání prací,
- zvláštní poznámku k odevzdávání,
- povolení odevzdávat práce i po stanoveném termínu odevzdání (ano/ne),
- generování výsledků kontroly ihned po odevzdání jedné práce nebo po uplynutí termínu odevzdání,
- vynechání uvedených bibliografických záznamů z nalezených podobností (ano/ne),
- vynechání citací v uvozovkách z kontroly (ano/ne),
- vynechání drobných nalezených shod (ano/ne – s uvedením délky v počtu slov nebo v % dokumentu),
- povolení zobrazení výsledků kontroly studentům (ano/ne),
- určení, jak bude uvedená práce zaindexována pro další kontroly (zakázáno, příp. označení repozitáře),

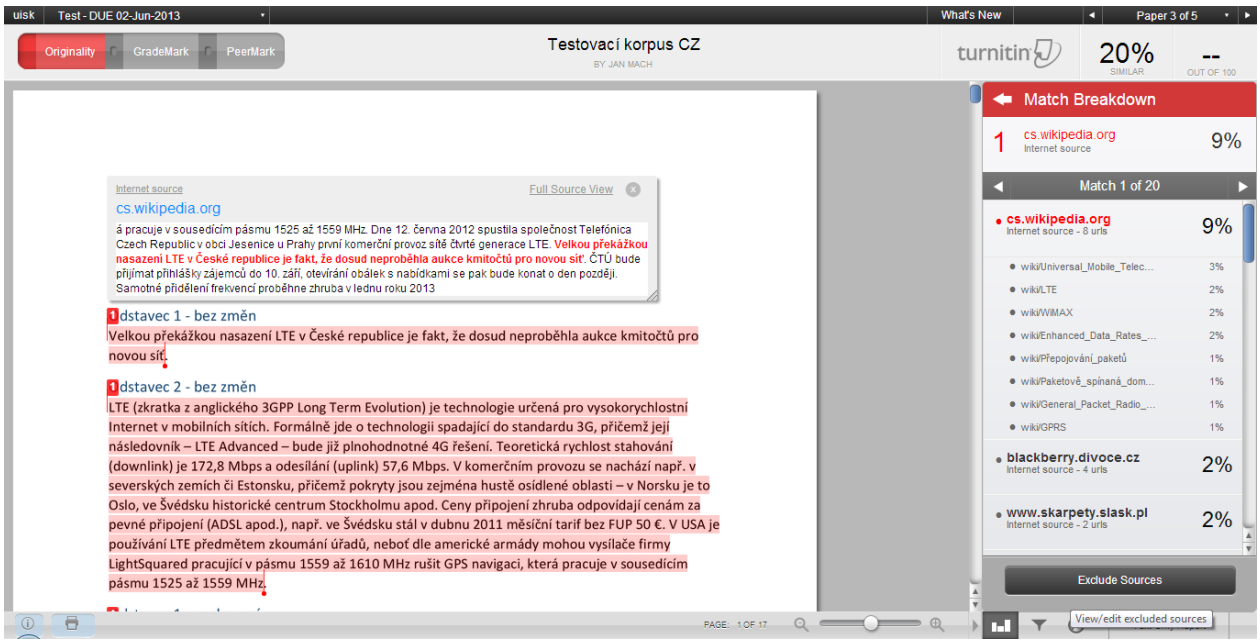
- určení, jaké zdroje prohledávat na podobnosti (repozitář studentských prací, současný a archivní Internet, periodika, časopisy a publikace).

Systém Turnitin umožňuje nahrávat soubory do velikosti 20 MB, maximální délka dokumentu je 400 stran. Podporované formáty jsou MS Word, WordPerfect, PostScript, PDF, HTML, RTF, OpenOffice (ODT), Hangul (HWP) a prostý text.

Kontrola duplicit

Samotné vyhledání duplicit dle provedených testů trvá okolo 30 sekund. Vyučující má pro každou odevzdávárnu k dispozici přehlednou tabulku se seznamem nahraných dokumentů s procentuálně, graficky a barevně znázorněnou škálou míry nalezených duplicit.

Samotné prohlížení nalezených duplicit je v rámci samostatného okna prohlížeče, kde levá část zobrazuje originální dokument s původním formátováním a s barevně vyznačenými podobnostmi. Pravá část obsahuje souhrn nalezených duplicit s uvedením hlavního zdroje a míry podobnosti. Pro každou duplicitu lze vypsat detailní přehled všech zdrojů, kde se daný text nachází. U každého zdroje lze zobrazit odpovídající duplicitní pasáž s textem indexovaným v systému Turnitin, příp. zobrazit zdroj online (viz Obrázek 22).



Obrázek 22 Rozhraní Turnitin pro kontrolu duplicit (zdroj: autor)

8.2.2 Ephorus

Základní údaje

Holandská aplikace Ephorus je určena pro kontrolu studentských prací, eVŠKP a dalších dokumentů. Práce na projektu Ephorus začala v roce 2003, v roce 2013 aplikaci používá přes 4000 škol a univerzit z celého světa (4 instituce v ČR), v Evropě má největší podíl na trhu (105).

Na rozdíl od aplikace Turnitin neobsahuje Ephorus již další, pokročilé funkce, jako je např. známkování nebo recenzování. Podle zdrojů provozovatele obsahuje databáze Ephorus miliardy webových stránek, dokumenty odeslané zapojenými školami a další zdroje jako jsou časopisy, referenční materiály aj. Přesné údaje o obsahu však provozovatel neuvádí a nelze tak pouze podle popisů aplikace důvěryhodně posoudit obsáhlost indexu aplikace.

Webové prostředí pro správu dat v aplikaci Ephorus obsahuje mimo jiných jazyků i české rozhraní. Vyučující může ve svém profilu nastavit:

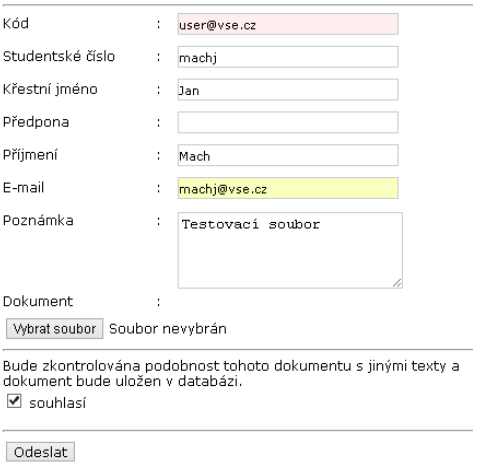
- údaje o účtu (jméno, heslo, e-mail, jazyk rozhraní),
- odevzdávací kódy ke svému účtu identifikující odevzdávárnu/kurz,
- zasílání e-mailu ihned, kdy student odevzdá práci,
- zasílání jednotlivých zpráv, pokud překročí dané procento podobnosti (např. při překročení 0 %, tj. posílat vždy),
- zasílání přehledu všech nových dokumentů za určité období (den, …).

Vyhledávání duplicit u eVŠKP je využíváno Fakultou podnikohospodářskou na Vysoké škole ekonomické v Praze. Popis funkcionality aplikace v této práci odpovídá konkrétní implementaci na této fakultě.

Způsob odevzdání

Pro studenty je k dispozici jednoduchý WWW formulář (viz Obrázek 23), kde kromě souboru základních metadat uvedou e-mail odpovědného vyučujícího. Po nahrání práce student získává, podobně jako v případě aplikace Turnitin, potvrzení o vložení práce do systému s identifikátorem vložené práce. Vyučující mohou využít pro nahrávání prací webové rozhraní aplikace. Nahrané práce není nijak možné ze systému smazat, proto při opakovaném nahrání bude předchozí práce vyhodnocena jako duplicitní.

Po vyhodnocení kontroly vyučující obdrží e-mail s výsledky kontroly, v případě testovacího korpusu od vložení práce do systému do odeslání výsledků e-mailem uplynula 1 hodina 45 minut.



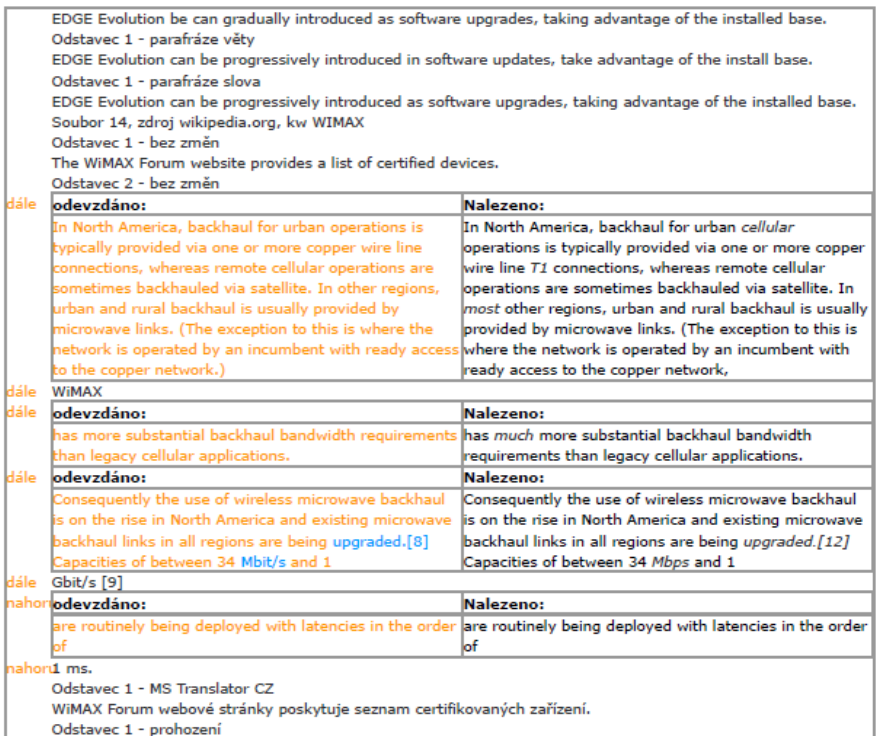
Obrázek 23 Vložení práce do systému Ephorus (zdroj: autor)

Kontrola duplicit

Výsledky kontroly jsou vyučujícímu zaslány e-mailem ve formě PDF příloh nebo jsou dostupné ve formátu HTML po přihlášení na web. Formátování zpráv je obdobné v PDF i na webu – z původního formátování je ponecháno zalomení odstavců, ale ne již formátování fontu jako v systému Turnitin.

E-mail obsahuje PDF se seznamem zdrojů, procentem podobnosti u každého zdroje a se zvýrazněným nalezeným textem v kontrolovaném dokumentu. Jednotlivé nalezené zdroje jsou v samostatných PDF souborech se zvýrazněným odpovídajícím textem v kontrolovaném a v nalezeném dokumentu (viz Obrázek 24).

Nalezené zdroje nejsou deduplikovány, tj. vyučující musí podobně jako u systémů MUNI pracně analyzovat všechny nalezené zdroje, i když některé odkazují na shodnou pasáž textu.


EDGE Evolution be can gradually introduced as software upgrades, taking advantage of the installed base.
Odstavec 1 - parafráze věty
EDGE Evolution can be progressively introduced in software updates, take advantage of the install base.
Odstavec 1 - parafráze slova
EDGE Evolution can be progressively introduced as software upgrades, taking advantage of the installed base.
Soubor 14, zdroj wikipedia.org, kw WIMAX
Odstavec 1 - bez změn
The WiMAX Forum website provides a list of certified devices.
Odstavec 2 - bez změn

dále	**odevzdáno:**	**Nalezeno:**
	In North America, backhaul for urban operations is typically provided via one or more copper wire line connections, whereas remote cellular operations are sometimes backhauled via satellite. In other regions, urban and rural backhaul is usually provided by microwave links. (The exception to this is where the network is operated by an incumbent with ready access to the copper network.)	In North America, backhaul for urban *cellular* operations is typically provided via one or more copper wire line *T1* connections, whereas remote cellular operations are sometimes backhauled via satellite. In *most* other regions, urban and rural backhaul is usually provided by microwave links. (The exception to this is where the network is operated by an incumbent with ready access to the copper network,
dále	WiMAX	
dále	**odevzdáno:**	**Nalezeno:**
	has more substantial backhaul bandwidth requirements than legacy cellular applications.	has *much* more substantial backhaul bandwidth requirements than legacy cellular applications.
dále	**odevzdáno:**	**Nalezeno:**
	Consequently the use of wireless microwave backhaul is on the rise in North America and existing microwave backhaul links in all regions are being upgraded.[8] Capacities of between 34 Mbit/s and 1	Consequently the use of wireless microwave backhaul is on the rise in North America and existing microwave backhaul links in all regions are being *upgraded.[12]* Capacities of between 34 *Mbps* and 1
dále	Gbit/s [9]	
nahoru	**odevzdáno:**	**Nalezeno:**
	are routinely being deployed with latencies in the order of	are routinely being deployed with latencies in the order of

nahoru 1 ms.
Odstavec 1 - MS Translator CZ
WiMAX Forum webové stránky poskytuje seznam certifikovaných zařízení.
Odstavec 1 - prohození

Obrázek 24 Zobrazení nalezených výsledků v systému Ephorus (zdroj: autor)

Drobná úprava textu jako je vypuštění či přidání jednoho slova je systémem detekována jako nepodstatná změna a systém zvýrazní celý text jako podobný (např. u zdrojů [C4] a [G4]). Ephorus má problém se správným parsováním vět, např. pasáž „upgraded.[8]“ u odstavce ze zdroje [B4] je interpretována jako jedno slovo a text proto není vyhodnocen jako shodný.

V testech nalezení duplicit dopadl systém Ephorus nejhůře. Nalezl pouze podobnost u 10 % zdrojů, a to jen především krátké, běžné fráze jako např. rozepsání zkraty WIMAX u zdroje [C4]. Převážně se nejednalo o nalezení delších pasáží textu, které by prokazovaly plagiátorství.

8.2.3 Systémy Masarykovy univerzity

Základní údaje

Systémy Masarykovy univerzity Theses.cz, Odevzdej.cz a Repozitar.cz kromě funkcí repozitáře eVŠKP, seminárních prací a odborných zaměstnaneckých děl slouží i jako nástroje

na vyhledávání duplicit. Přes specializaci na různé typy textů mají aplikace na pozadí stejný algoritmus a databázi, vůči které se duplicity vyhledávají. V práci se zaměřujeme na aplikaci Theses.cz, která slouží jako národní registr VŠKP a systém na kontrolu plagiátorství.

Aplikace byly vyvinuty v rámci rozvojových centralizovaných projektů MŠMT ČR a provozuje je Fakulta informatiky Masarykovy univerzity. Do roku 2012 byly tyto aplikace pro školy účastnící se projektů zdarma, od roku 2013 je vyhledávání duplicit zpoplatněno v částce řádově desítek tisíc Kč ročně na univerzitu, podle počtu studentů. Nově bylo také zpoplatněno zodpovídání dotazů studentů u veřejného rozhraní aplikace Odevzdej.cz (seminární práce).

Popsaná funkcionalita odpovídá verzi Theses.cz v době konání testu, tj. v roce 2013.

Způsob odevzdání

Aplikace Theses.cz podporuje několik možností, jak práce ke kontrole do systému zavést. Autorizovaní studenti mohou po přihlášení na web Theses.cz vyplnit formulář a nahrát práci přímo, autorizovaní administrátoři mohou kromě vložení jedné práce formulářem použít hromadné zavedení prací pomocí souboru XML. Je možné použít buď národní standard EVSKP-MS (1) připravený Odbornou komisí pro otázky elektronického zpřístupňování vysokoškolských kvalifikačních prací AKVŠ ČR nebo proprietární formát Theses.cz. Oba formáty jsou obsahově obdobné (viz kapitola 4 *Mapování metadat eVŠKP*), obsahují metadata popisující danou práci a URL adresy, ze kterých systém následně stáhne a zpracuje odpovídající soubory eVŠKP (hlavní práce, přílohy, posudky).

Pro hromadný import pomocí XML je možné zvolit buď nahrání souboru přes web Theses.cz s omezením na 5000 záznamů na jeden XML soubor, nebo automatizovaný přístup přes protokol OAI-PMH využitý např. Vysokou školou ekonomickou v Praze.

Kontrola duplicit

U systémů MUNI je výsledek kontroly dostupný v řádu několika hodin po nahrání. Vzhledem k takto velké době zpracování byl proveden test na veřejném rozhraní Odevzdej.cz, kde libovolná osoba může nahrát práci ke kontrole a výsledek je po zpracování uživateli zaslán e-mailem. Testovaný soubor byl aplikací Odevzdej.cz zpracován za 12 hodin a 4 minuty.

Podobné časy na kontrolu jsou i u systému Theses.cz. Kontrola dokumentů proto u systémů MUNI není možná ihned po nahrání, ale většinou až druhý den.

Antiplagiátorské systémy MUNI pracují na principu porovnávání páru dokumentů vůči sobě. Analyzovaný dokument je tedy porovnán postupně se všemi indexovanými dokumenty a aplikace pak zobrazí zprávu pro každou takovouto analyzovanou dvojici samostatně. V době konání testu tedy nebylo možné zobrazení všech nalezených duplicit v analyzovaném dokumentu najednou, pouze po částech podle jednotlivých nalezených dokumentech. Systém neuměl spočítat celkové procento podobnosti analyzovaného dokumentu vůči všem nalezeným dokumentům, tj. zobrazoval pouze procentuální podobnost jednotlivých dvojic dokumentů.

Systémy MUNI ve verzi použité v testu informovaly pouze o podobnosti delší než 5 % z jednoho dokumentu v porovnávaném páru. Pokud by plagiát obsahoval celkem 100 stránek, systém by nenahlásil až 4 strany textu opsaného z jiného stejně velkého dokumentu. V důsledku toho, kdyby např. stostránkový dokument byl složen ze zkopírovaných čtyřstránkových pasáží, každá z jiného stostránkového dokumentu, nemuselo by být přesto plagiátorství detekováno.

Z uvedeného omezení na min. 5 % podobnosti vyplývá závažné riziko, že rozsáhlá část zkopírovaného textu nebude detekována, pokud bude tvořit do 5 % plagiátu (tj. minimálně 2 až 3 strany u průměrné VŠKP) a zároveň do 5 % originálního souboru.

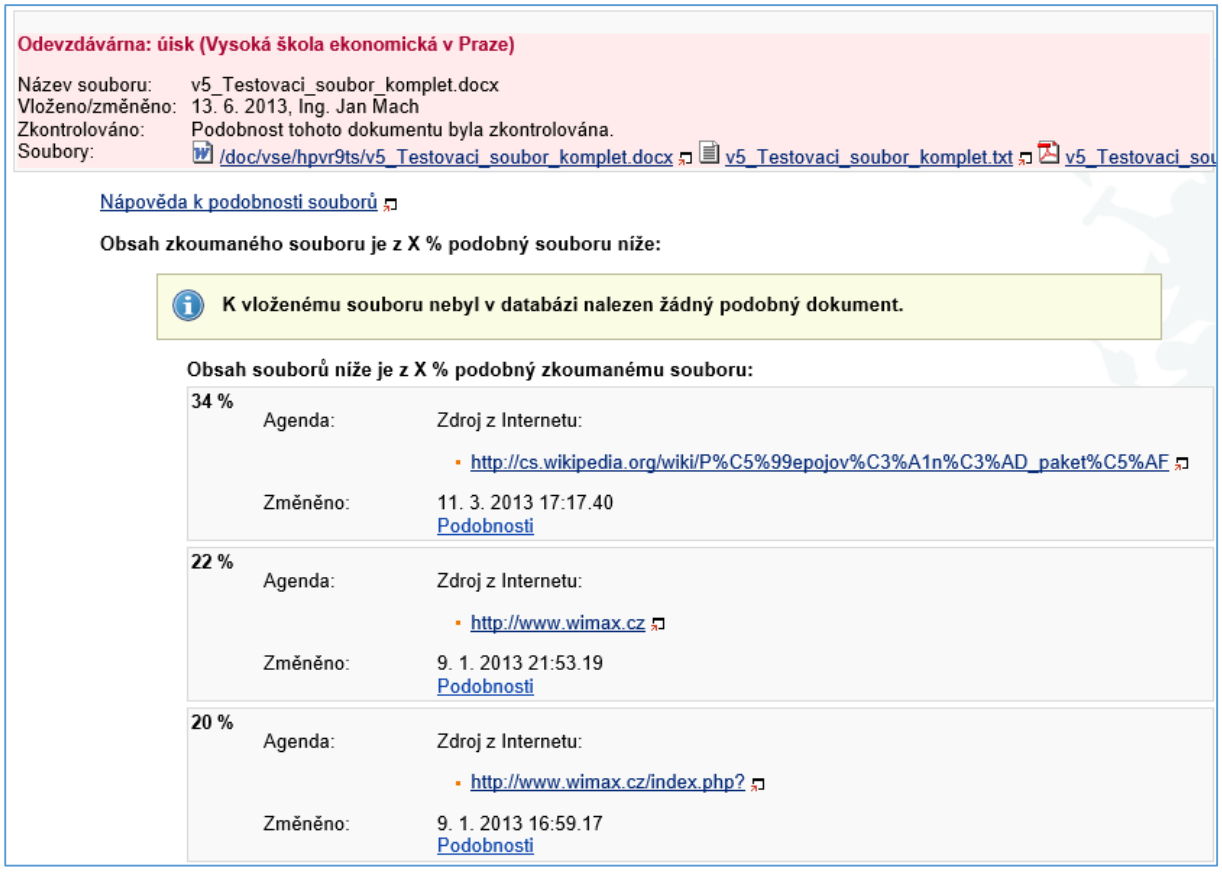
Vzhledem k tomu, že dochází k analýze dokumentů po párech, systémy MUNI ve vyhodnocení zobrazovaly pro analýzu dva seznamy (viz Obrázek 25):

a) seznam dokumentů a podobností, u kterých shodný text tvoří min. 5 % nového, zkoumaného dokumentu (označení v Theses.cz „Obsah zkoumaného souboru je z X % podobný souboru níže:“),
b) seznam dokumentů a podobností, u kterých shodný text tvoří min. 5 % staršího, porovnávaného dokumentu z indexu (označení v Theses.cz „Obsah souborů níže je z X % podobný zkoumanému souboru:“)

Komplikovaný dvojí pohled na podobnosti přislíbil tým Masarykovy univerzity odstranit v nové verzi aplikace, k čemuž došlo na podzim 2014.

Z důvodu omezení 5% hranicí a dualitě pohledu na podobnosti došlo u analyzovaného dokumentu s testovacím korpusem k následujícím výsledkům (viz Obrázek 25):

1) V seznamu „Obsah zkoumaného souboru je z X % podobný souboru níže:“ systém nenalezl žádnou podobnost, závěr: „K vloženému souboru nebyl v databázi nalezen žádný podobný dokument.“.
2) V seznamu „Obsah souborů níže je z X % podobný zkoumanému souboru:“ bylo nalezeno 20 dokumentů s podobností 6 – 34 % z originálního textu, z toho většina dokumentů odkazovala na shodnou pasáž v testovacím souboru.



Obrázek 25 Vyhodnocení podobností systémy MUNI (zdroj: autor)

Nalezené podobnosti testovaného a původního dokumentu jsou vyučujícím dostupné v textovém výpisu, kdy systém zachovává pouze zalomení odstavců. Velikost písmen, nadpisů apod. formátování je ignorováno a orientace v textu je tak složitější než v případě systému Turnitin.

V databázi systémů MUNI existuje větší množství dokumentů zavedených duplicitně, mj. shodné studijní materiály uložené do databáze v rozmezí pár sekund. Např. velké

množství dokumentů nalezených v testu u zdroje [A5] odkazuje na shodnou pasáž s definicí Wi-Fi podle české Wikipedie. Vyučující tak musí projít všechny nalezené dvojice a jejich podobnosti i v případě, kdy různé dokumenty ukazují na stejný text.

Systémy MUNI by měly mít velkou výhodu v detekci českých vysokoškolských kvalifikačních prací a seminárních prací, které v sobě mají indexovány. Bohužel, pravděpodobně vzhledem k podmínce detekce min. 5 %, aplikace našla jen jednu z deseti kvalifikačních prací v testovacím korpusu, a to ještě jen jako podobnost vůči externímu dokumentu z webu se shodnou pasáží. Přitom samotné vyhledávácí rozhraní Theses.cz dokáže podobné eVŠKP ve své databázi najít. Dokud nebude upravena hranice pro vyhodnocení duplicit pod 5 % podobnosti, přínos systémů MUNI pro hledání duplicit ve vysokoškolských kvalifikačních pracích je tak sporný.

Systémy MUNI nemají problém jako duplicitní vyhodnotit věty bez diakritiky. V případě prohození či záměny slov ve větě jsou většinou uvedená slova přeskočena a jako duplicitní je uveden pouze okolní text. Někdy však analýza dané věty selže a podobnost kratšího textu není nalezena (např. při chybějící čárce ve větě, zdroj [B1]).

Zásadní nevýhody systémů MUNI v době konání testu byly:

a) omezení na min. detekovanou podobnost 5 %,
b) dualita pohledu na podobnosti,
c) duplicita zdrojů.

8.2.4 GooglePlagiarism

Základní údaje

Desktopovou aplikaci GooglePlagiarism pro operační systém Windows autor této práce naprogramoval v březnu 2010, později dle potřeby upravoval do současné podoby použité pro tento test. Aplikace byla využita např. pro přípravu podkladů analýz autora disertační práce pro pořady České televize Reportéři ČT a Události, komentáře (106, 107, 108).

Aplikace na základě syntaktické analýzy dělí (parsuje) vstupní dokument Microsoft Word na věty, které následně přes webové rozhraní Web Search API vyhledávače Google vyhledává na přesnou shodu, tj. jako frázi.

Vzhledem k omezení služby Google na množství dotazů v kratším časovém intervalu je v aplikaci dodatečně doprogramován mechanismus časových prodlev (v řádu několika minut mezi dotazy) v případě aktivace blokování ze strany Google. Limit na množství dotazů by šel odstranit, a tím i významně snížit celkový čas na analýzu, využitím placeného vyhledávání přes Google Custom Search API. V únoru 2015 činí cena za 1 000 dotazů 5 USD, 100 dotazů denně zdarma, limit max. 10 000 dotazů denně (109).

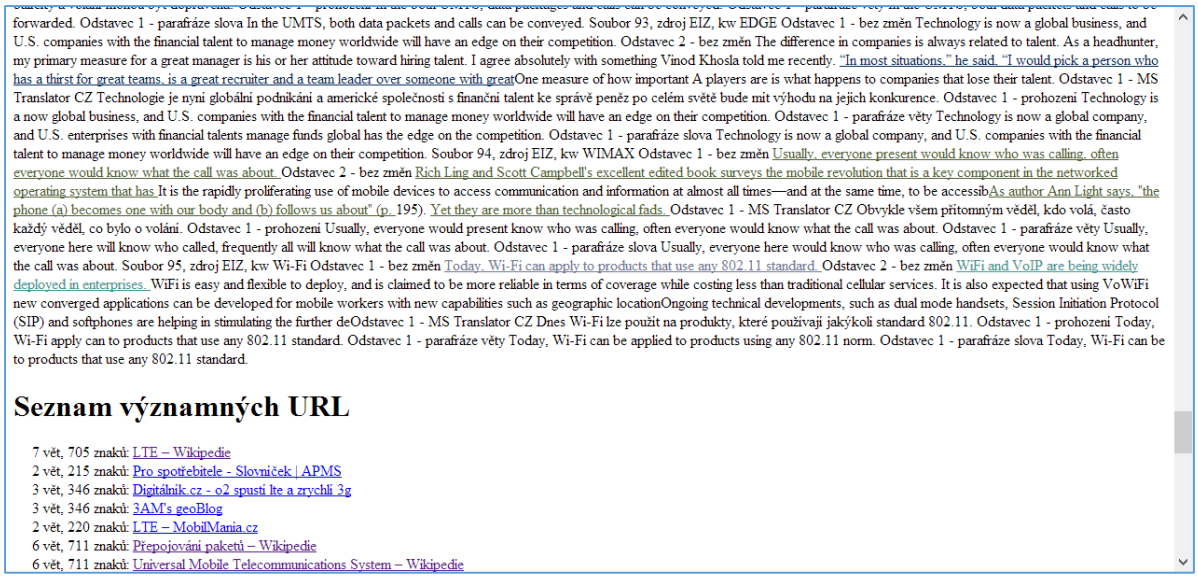
Prodlevy během jednotlivých vyhledávání zvětšují potřebnou dobu na provedení analýzy, v případě testovacího korpusu trvalo zpracování 3 hodiny. Tento čas je však stále menší, než kolik na analýzu potřebují systémy MUNI.

Způsob odevzdání

Oproti výše uvedeným systémům zaměřeným na centralizované zpracování výsledků, GooglePlagiarism je uživatelskou aplikací určenou k provozu na lokálním PC. Aplikace neřeší proces odevzdávání prací, vyučující pouze v aplikaci vybírá lokální dokument určený ke kontrole.

Kontrola duplicit

Výsledky analýzy jsou dostupné ve formátu HTML zprávy s barevně zvýrazněnými nalezenými originálními zdroji, odkazy a s přehledem nalezených zdrojů. Odstranění veškerého formátování primárního dokumentu výrazně snižuje orientaci ve vygenerované zprávě. Formát HTML s odkazy na nalezené zdroje přímo v textu zvyšuje komfort prohlížení zdrojů. Ukázka výsledné analýzy je na Obrázku 26.


forwarded. Odstavec 1 - parafráze slova In the UMTS, both data packets and calls can be conveyed. Soubor 93, zdroj EIZ, kw EDGE Odstavec 1 - bez změn Technology is now a global business, and U.S. companies with the financial talent to manage money worldwide will have an edge on their competition. Odstavec 2 - bez změn The difference in companies is always related to talent. As a headhunter, my primary measure for a great manager is his or her attitude toward hiring talent. I agree absolutely with something Vinod Khosla told me recently. "In most situations," he said, "I would pick a person who has a thirst for great teams, is a great recruiter and a team leader over someone with greatOne measure of how important A players are is what happens to companies that lose their talent. Odstavec 1 - MS Translator CZ Technologie je nyní globální podnikání a americké společnosti s finanční talent ke správě peněz po celém světě bude mít výhodu na jejich konkurence. Odstavec 1 - prohození Technology is a now global business, and U.S. companies with the financial talent to manage money worldwide will have an edge on their competition. Odstavec 1 - parafráze věty Technology is now a global company, and U.S. enterprises with financial talents manage funds global has the edge on the competition. Odstavec 1 - parafráze slova Technology is now a global company, and U.S. companies with the financial talent to manage money worldwide will have an edge on their competition. Soubor 94, zdroj EIZ, kw WIMAX Odstavec 1 - bez změn Usually, everyone present would know who was calling, often everyone would know what the call was about. Odstavec 2 - bez změn Rich Ling and Scott Campbell's excellent edited book surveys the mobile revolution that is a key component in the networked operating system that has It is the rapidly proliferating use of mobile devices to access communication and information at almost all times—and at the same time, to be accessibAs author Ann Light says, "the phone (a) becomes one with our body and (b) follows us about" (p. 195). Yet they are more than technological fads. Odstavec 1 - MS Translator CZ Obvykle všem přítomným věděl, kdo volá, často každý věděl, co bylo o volání. Odstavec 1 - prohození Usually, everyone would present know who was calling, often everyone would know what the call was about. Odstavec 1 - parafráze věty Usually, everyone here will know who called, frequently all will know what the call was about. Odstavec 1 - parafráze slova Usually, everyone here would know who was calling, often everyone would know what the call was about. Soubor 95, zdroj EIZ, kw Wi-Fi Odstavec 1 - bez změn Today, Wi-Fi can apply to products that use any 802.11 standard. Odstavec 2 - bez změn WiFi and VoIP are being widely deployed in enterprises. WiFi is easy and flexible to deploy, and is claimed to be more reliable in terms of coverage while costing less than traditional cellular services. It is also expected that using VoWiFi new converged applications can be developed for mobile workers with new capabilities such as geographic locationOngoing technical developments, such as dual mode handsets, Session Initiation Protocol (SIP) and softphones are helping in stimulating the further deOdstavec 1 - MS Translator CZ Dnes Wi-Fi lze použít na produkty, které používají jakýkoli standard 802.11. Odstavec 1 - prohození Today, Wi-Fi apply can to products that use any 802.11 standard. Odstavec 1 - parafráze věty Today, Wi-Fi can be applied to products using any 802.11 norm. Odstavec 1 - parafráze slova Today, Wi-Fi can be to products that use any 802.11 standard.

Seznam významných URL

7 vět, 705 znaků: LTE – Wikipedie
2 vět, 215 znaků: Pro spotřebitele - Slovníček | APMS
3 vět, 346 znaků: Digitálník.cz - o2 spustí lte a zrychlí 3g
3 vět, 346 znaků: 3AM's geoBlog
2 vět, 220 znaků: LTE – MobilMania.cz
6 vět, 711 znaků: Přepojování paketů – Wikipedie
6 vět, 711 znaků: Universal Mobile Telecommunications System – Wikipedie

Obrázek 26 Výsledek analýzy GooglePlagiarism (zdroj: autor)

Vzhledem k nevhodné implementaci parsování vět v aplikačním programovém rozhraní Microsoft Word bylo pro hledání občas použito vět přerušených např. v půlce slova nebo chybně ukončených při chybné interpretaci tečky u zkratky ve větě. Vyhledávání na přesnou shodu v Google pak uvedené věty nemůže najít, ve výsledku je detekována jen část z celého fragmentu opsaného textu. I částečná detekce je sice jako indikátor postačující, přesto by bylo vhodné v aplikaci implementovat pokročilejší parsování vět zohledňující lépe tečky u zkratek, závorky, horní indexy apod.

Vyhledávač Google měl občas problém najít frázi bez diakritiky, i když frázi s diakritikou našel. Důvodem je použití uvozovek pro vyhledávání fráze, tj. textu bez změny. Pokud bychom však použili volné vyhledávání (bez uvozovek), zvýšil by se významně počet falešně pozitivních nálezů.

Další problémy byly zjištěny při vyhledávání dokumentů zaindexovaných anglickou verzí Google http://www.google.com (např. zdroj [J3]), kdy programové rozhraní vyhledávače Google dokázalo najít pouze 3 z 5 dokumentů (z toho přesná shoda vět 3×, přesná shoda u odstavce jen 1×). I přes to se v testu jednalo o nejlepší výsledek vyhledávání duplicit vůči zdrojům z google.com.

Další nepřesnosti mohou způsobit zdrojové dokumenty v PDF (např. zdroj [J5]), kdy Google špatně identifikuje tok textu při zalomení řádku, nedokáže extrahovat větu jako celek a má tak problémy dlouhé fráze nalézt.

I přes problémy s parsováním vět aplikace GooglePlagiarism nalezla shody (přinejmenším částečné) u 58 % testovaných dokumentů. Dosáhla tak nejlepšího skóre v celkovém množství nalezených dokumentů z testovaných aplikací i přes to, že nemá přístup ke všem plným textům vysokoškolských kvalifikačních prací a EIZ.

Zpřesnění detekce by bylo možné díky implementaci pokročilejší parsovací funkce a vyhledávání po kratších úsecích než věta, neboť oproti konkurenci při vyhledávání celých vět jako frází je problém nalézt zdrojový dokument i v případě jen malé změny věty (parafráze, vypuštění slova).

Lepší orientaci ve vygenerované zprávě by posloužilo alespoň částečné zachování formátování zdrojového dokumentu. Snížení celkové doby vyhledávání lze dosáhnout pomocí následujících metod:

- placené vyhledávání s vyšším limitem,
- paralelizace vyhledávání z více počítačů,
- ukládání již nalezených originálních dokumentů na straně klienta do vyrovnávací paměti, parsování obsahu, indexace a primární vyhledávání textu v takovémto lokálním indexu.

8.3 Výsledky testů

Připravený textový korpus byl vyhodnocen jednotlivými testovanými systémy na přítomnost duplicit, nálezy byly ručně posouzeny a vyhodnoceny jako:

a) *Nález přímo prokazující plagiátorství* obsahující signifikantní úsek textu jak v originálním textu, tak v korpusu (plagiátu). Označení v Příloze XI znakem „+“.
b) *Nález zakládající podezření na plagiátorství*. Např. nalezení velmi krátké fráze v obsahově odlišném zdrojovém dokumentu. Označení v Příloze XI znakem „o“.
c) *Nález náhodný*, nezakládající podezření na plagiátorství. Nález velmi obecné, typicky velmi často používané skupiny slov nebyl započítán jako relevantní, v Příloze XI pole bez označení.

Kompletní výsledky testů jednotlivých systémů jsou uvedeny v Příloze XI a online (DOI 10.13140/2.1.4398.4162), jednotlivé faktory pro hodnocení jsou detailněji zpracovány do samostatných tabulek níže. Pro každé hodnocení je uvedena tabulka s počtem nálezů v celých

číslech (zvýraznění počtu nálezů barevnou škálou – zelená barva označuje úspěšnější systém) a tabulka s procentuálním podílem (zvýraznění míry úspěšnosti v procentech datovým pruhem).

8.3.1 Použité zkratky v tabulkách

Tabulka 5 obsahuje seznam zkratek použitých v záhlaví tabulek následujících v této kapitole.

Tabulka 5 Použité zkratky v tabulkách (zdroj: autor)

Zkratka	Vysvětlivka
PLG	Textový korpus použitý k testování (100% plagiát)
THE	Systém MUNI Theses.cz
TUR	Turnitin
EPH	Ephorus
GPL	GooglePlagiarism – aplikace autora

8.3.2 Počet nalezených záznamů podle zdroje

Tabulka 6 a Tabulka 7 znázorňují úspěšnost vyhledávání duplicit v jednotlivých zdrojích. Každý zdroj byl v testovacím korpusu zastoupen pěti záznamy, celkem 50 záznamů.

Tabulka 6 ukazuje nulovou úspěšnost systémů při vyhledávání duplicit z mediálních zdrojů monitorovaných v aplikaci Anopress, kde plné texty nejsou na Internetu veřejně dostupné.

Tabulka 7 ukazuje významnou převahu úspěšnosti detekce duplicit u systémů Turnitin (celkem 44 %) a GooglePlagiarism (58 %), naopak nízkou úspěšnost Theses.cz (14 %) a především Ephorus (10 %). Nejlépe dohledatelné jsou zdroje nalezené českou verzí Google a z Wikipedie.

Tabulka 6 Počet nalezených záznamů podle zdroje 1 (zdroj: vlastní zpracování)

Kategorie	PLG	THE	TUR	EPH	GPL	Průměr
wikipedia.cz	5	3	5	2	5	3,75
wikipedia.org (en)	5	1	3	2	5	2,75
VŠKP (cz)	5	1	2	1	1	1,25
VŠKP (en)	5	0	3	0	2	1,25
NDLTD	5	0	0	0	1	0,25
Anopress	5	0	0	0	0	0
Arxive.org	5	0	1	0	3	1
Google.cz (cz)	5	2	3	0	5	2,5
Google.com (en)	5	0	2	0	3	1,25
EIZ	5	0	3	0	4	1,75
Celkem	50	7	22	5	29	15,75

Tabulka 7 Počet nalezených záznamů podle zdroje 2 (zdroj: vlastní zpracování)

Kategorie	PLG	THE	TUR	EPH	GPL	Průměr
wikipedia.cz	100%	60%	100%	40%	100%	75%
wikipedia.org (en)	100%	20%	60%	40%	100%	55%
VŠKP (cz)	100%	20%	40%	20%	20%	25%
VŠKP (en)	100%	0%	60%	0%	40%	25%
NDLTD	100%	0%	0%	0%	20%	5%
Anopress	100%	0%	0%	0%	0%	0%
Arxive.org	100%	0%	20%	0%	60%	20%
Google.cz (cz)	100%	40%	60%	0%	100%	50%
Google.com (en)	100%	0%	40%	0%	60%	25%
EIZ	100%	0%	60%	0%	80%	35%
Průměr	100%	14%	44%	10%	58%	32%

8.3.3 Počet nalezených záznamů podle formátu

Hodnocení vyhledávání duplicit podle typu dokumentu neukazuje dle tabulek 8 a 9 zásadní rozdíl mezi zdrojem v HTML či v PDF. V testovacím korpusu byl pouze jeden dokument ve formátu DOC (MS Word), nalezený aplikací GooglePlagiarism.

Tabulka 8 Počet nalezených záznamů podle formátu 1 (zdroj: vlastní zpracování)

Formát	PLG	THE	TUR	EPH	GPL	Průměr
PDF	24	1	9	1	10	5,25
HTML	25	6	13	4	18	10,25
doc	1	0	0	0	1	0,25
Celkem	50	7	22	5	29	15,75

Tabulka 9 Počet nalezených záznamů podle formátu 2 (zdroj: vlastní zpracování)

Formát	PLG	THE	TUR	EPH	GPL	Průměr
PDF	100%	4%	38%	4%	42%	38%
HTML	100%	24%	52%	16%	72%	53%
doc	100%	0%	0%	0%	100%	40%

8.3.4 Počet nalezených záznamů podle jazyka

Podle tabulek 10 a 11 má pouze systém Theses.cz nižší úspěšnost při vyhledávání duplicit z anglických dokumentů, ostatní vykazují vyrovnanější výsledky mezi českými a anglickými zdroji. V testovacím korpusu byl jeden dokument slovensky, nebyl nalezen žádným systémem (zdroj [F4] - Anopress).

Tabulka 10 Počet nalezených záznamů podle jazyka 1 (zdroj: vlastní zpracování)

Jazyk	PLG	THE	TUR	EPH	GPL	Průměr
česky	19	6	10	3	11	7,5
anglicky	30	1	12	2	18	8,25
slovensky	1	0	0	0	0	0
Celkem	50	7	22	5	29	15,75

Tabulka 11 Počet nalezených záznamů podle jazyka 2 (zdroj: vlastní zpracování)

Jazyk	PLG	THE	TUR	EPH	GPL	Průměr
česky	100%	32%	53%	16%	58%	39%
anglicky	100%	3%	40%	7%	60%	28%
slovensky	100%	0%	0%	0%	0%	0%

8.3.5 Počet nalezených záznamů podle data publikování

V tabulkách 12 a 13 je znázorněna úspěšnost vyhledávání podle roku publikování zdrojového dokumentu (pokud nebylo známo datum publikování, byl přiřazen aktuální rok testu 2013).

Tabulka 12 Počet nalezených záznamů podle data publikování 1 (zdroj: vlastní zpracování)

Rok	PLG	THE	TUR	EPH	GPL	Průměr
2013	24	6	12	4	17	9,75
2012	10	0	3	0	3	1,5
starší	16	1	7	1	9	4,5
Celkem	50	7	22	5	29	15,75

Tabulka 13 Počet nalezených záznamů podle data publikování 2 (zdroj: vlastní zpracování)

Rok	PLG	THE	TUR	EPH	GPL	Průměr
2013	100%	25%	50%	17%	71%	41%
2012	100%	0%	30%	0%	30%	15%
starší	100%	6%	44%	6%	56%	28%

8.3.6 Počet všech nalezených záznamů podle typu úprav

Hodnocení v tabulkách 14 a 15 ukazují úspěšnost vyhledávání podle jednotlivých typů úprav textu ze zdrojového dokumentu (viz oddíl 8.1.3 *Použité transformace*).

Z tabulky 14 je vidět téměř nulová schopnost aplikací vyhledávat překlady originálních textů. Tabulka 15 jasně ukazuje velmi nízkou úroveň detekce jednotlivých úprav systémy Ephorus a GooglePlagiarism. Naopak systém Turnitin vykazuje vyrovnanou úspěšnost i při různých úpravách zdrojové věty.

Tabulka 14 Počet všech nalezených záznamů podle typu úprav 1 (zdroj: vlastní zpracování)

Úprava	PLG	THE	TUR	EPH	GPL	Průměr
jedna věta	**50**	**6**	**20**	**1**	**28**	**13,75**
jeden odstavec	**50**	**7**	**21**	**3**	**23**	**13,5**
prohození slova	**50**	**6**	**20**	**1**	**0**	**6,75**
bez diakritiky	**19**	**5**	**9**	**1**	**8**	**5,75**
parafráze věty	**31**	**0**	**10**	**0**	**0**	**2,5**
parafráze slova	**50**	**4**	**20**	**1**	**1**	**6,5**
překlad	**50**	**0**	**0**	**1**	**0**	**0,25**
Celkem	**300**	**28**	**100**	**8**	**60**	**49**

Tabulka 15 Počet všech nalezených záznamů podle typu úprav 2 (zdroj: vlastní zpracování)

Úprava	PLG	THE	TUR	EPH	GPL	Průměr
jedna věta	100%	12%	40%	2%	56%	28%
jeden odstavec	100%	14%	42%	6%	46%	27%
prohození slova	100%	12%	40%	2%	0%	14%
bez diakritiky	100%	26%	47%	5%	42%	30%
parafráze věty	100%	0%	32%	0%	0%	8%
parafráze slova	100%	8%	40%	2%	2%	13%
překlad	100%	0%	0%	2%	0%	1%
Průměr	**100%**	**10%**	**35%**	**3%**	**21%**	**17%**

8.3.7 Počet přesně nalezených záznamů podle typu úprav

Tabulky 16 a 17 jsou obdobou tabulek 14 a 15, ale jsou započítány pouze nálezy takových zdrojů, kdy je jasně možné dokázat plagiátorství. Z porovnání tabulek vyplývá, že systém

Ephorus jen v jednom případě nalezl odpovídající zdrojový dokument, a to v případě hledání celého odstavce textu. Ostatní nálezy započítané v tabulkách 14 a 15 u aplikace Ephorus jsou obecné fráze.

Z tabulek 16 a 17 je zřejmé, že testované systémy nejsou schopny nalézt překlady originálních textů. Z průzkumu (18) (viz 2.1.5 Definice plagiátorství) však víme, že tato metoda plagiátorství vyžadující náročnější práci plagiátora není mezi nejčastěji používanými.

Tabulka 16 Počet přesně nalezených záznamů podle typu úprav 1 (zdroj: vlastní zpracování)

Úprava	PLG	THE	TUR	EPH	GPL	Průměr
jedna věta	50	5	8	0	25	9,5
jeden odstavec	50	6	10	1	9	6,5
prohození slova	50	1	7	0	0	2
bez diakritiky	19	4	6	0	7	4,25
parafráze věty	31	0	2	0	0	0,5
parafráze slova	50	3	8	0	1	3
překlad	50	0	0	0	0	0
Celkem	**300**	**19**	**41**	**1**	**42**	**25,75**

Tabulka 17 Počet přesně nalezených záznamů podle typu úprav 2 (zdroj: vlastní zpracování)

Úprava	PLG	THE	TUR	EPH	GPL	Průměr
jedna věta	100%	10%	16%	0%	50%	19%
jeden odstavec	100%	12%	20%	2%	18%	13%
prohození slova	100%	2%	14%	0%	0%	4%
bez diakritiky	100%	21%	32%	0%	37%	22%
parafráze věty	100%	0%	6%	0%	0%	2%
parafráze slova	100%	6%	16%	0%	2%	6%
překlad	100%	0%	0%	0%	0%	0%
Průměr	**100%**	**7%**	**15%**	**0%**	**15%**	**9%**

8.3.8 Hodnocení ovládání a funkcí systémů

Následující tabulka 18 zobrazuje autorovo subjektivní hodnocení ovládání a funkcí antiplagiátorských systémů. Detailní popis nástrojů byl uveden v podkapitole 8.2 *Zhodnocení antiplagiátorských systémů*. Jednotlivá hodnocení a jejich kritéria jsou:

- **doba zpracování** – doba od vložení dokumentu po obdržení výsledku,
- **přehlednost výsledků** – zachování formátování, srozumitelnost nalezených podobností,

- **zobrazení celkové podobnosti** – schopnost systémů zobrazit původní dokument se všemi nalezenými podobnostmi, zobrazení procentuální míry podobnosti analyzovaného dokumentu se všemi dalšími zdroji,
- **minimální podobnost** – schopnost ovlivnit minimální délku fráze (např. počet slov), která bude vyhledávána jako celek/hlášena jako nalezená podobnost,
- **cena** – celkové náklady na provoz aplikace,
- **integrace s IS školy** – dostupnost aplikačního programového rozhraní (API) pro napojení na interní informační systém školy, míra integrace v testovacím prostředí,
- **deduplikace zdrojů** – agregované zobrazení všech zdrojů nalezených pro jednu pasáž analyzovaného dokumentu, bez nutnosti procházení všech výsledků individuálně.

Z tabulky vyplývá zásadní rozdíl mezi systémy Theses.cz a Turnitin, kdy systém Theses.cz vyniká nízkou cenou a velmi dobrým aplikačním rozhraním pro integraci s informačním systémem školy. Naopak Theses.cz velmi zaostával díky dlouhé době pro zpracování dokumentů, malou přehledností výsledků, absencí deduplikace zdrojů a neschopností spočítat celkovou podobnost (v testu zobrazoval pouze dvojice dokumentů bez zachování formátování, neuměl spočítat celkové procento podobnosti). Velkou nevýhodou systému Theses.cz byla také podmínka nalezení min. 5% podobnosti z celého dokumentu pro zobrazení podobnosti oproti možnosti variabilního nastavení počtu shodných slov v systému Turnitin. Tato omezující podmínka byla v systému Theses.cz odstraněna v roce 2014, po publikování výsledků tohoto testu.

Systém Ephorus je v hodnocení funkčnosti ovládání na pomyslném středu, jeho užitečnost však sráží špatné výsledky vyhledávání okomentované v předcházejících kapitolách. Aplikace GooglePlagiarism je primárně určena pro osobní použití, v případě placené verze vyhledávání Google by výrazně klesla doba potřebná na zpracování výsledků na úkor ztráty současné bezplatnosti vyhledávání (při placeném vyhledávání by byla cena řádově 100 Kč/VŠKP, dle počtu dotazů).

Tabulka 18 Hodnocení ovládání a funkcí systémů (zdroj: vlastní zpracování)

Hodnocení	THE	TUR	EPH	GPL
doba zpracování	✗	✓	!	✗
přehlednost výsledků	✗	✓	!	!
zobrazení celkové podobnosti	✗	✓	!	!
minimální podobnost	✗	✓	✓	!
cena	✓	✗	!	✓
integrace s IS školy	✓	✗	!	✗
deduplikace zdrojů	✗	✓	✗	✓

8.4 Vyhodnocení hypotéz

Úkolem testu bylo ověřit následující hypotézy o úspěšnosti detekce duplicit testovanými systémy:

1. Aplikace umí odhalit jednu větu zkopírovanou ze zdrojového dokumentu.
2. Aplikace umí odhalit jeden odstavec zkopírovaný ze zdrojového dokumentu. Aplikaci nevadí případná zalomení řádků, indexy apod. ve zdrojovém nebo testovaném dokumentu.
3. Pro úspěšnou detekci nevadí, pokud plagiátor přidá/odebere slovo v kopírované větě.
4. Aplikace provádí detekci českých textů nezávisle na diakritice.
5. Pro úspěšnou detekci nevadí, pokud plagiátor parafrázuje jedno slovo ve větě.
6. Pro úspěšnou detekci nevadí, pokud plagiátor parafrázuje celou větu.
7. Pro úspěšnou detekci nevadí, pokud plagiátor přeloží text z/do českého jazyka.

U jednotlivých zdrojů plagiátorství jsme si dále stanovili tyto hypotézy o zdrojích a o chování antiplagiátorských systémů:

8. Systém Theses.cz by měl v detekci plagiátorství u českých eVŠKP dosahovat nejlepších výsledků.
9. U zdroje Anopress bude nízké procento nalezených podobností oproti zdrojům volně dostupným na Internetu.
10. Lepších výsledků u elektronických informačních zdrojů a Open Access zdrojů dosáhnou nástroje zahraniční oproti českým.
11. Velmi dobrých výsledků u webových zdrojů budou dosahovat systémy využívající služeb webových vyhledávačů.

Při vyhodnocení výše uvedených hypotéz 1-7 vycházíme z 300 vzorků textu dle tabulky 16, která obsahuje počty všech nalezených záznamů podle typu úprav. U hodnocení hypotéz 8 až 12 vycházíme z tabulky 6 Počet nalezených záznamů podle zdroje 1.

Na základě výsledků testu můžeme stanovit předpokládanou platnost hypotéz 1 a 2 jako podíl nalezených záznamů pro větu/odstavec v daném systému k počtu záznamů v celém textovém korpusu (tj. 50 v tomto testu). V případě hypotéz 3-7 stanovíme platnost hypotéz u daného systému jako podíl počtu nalezeného typu úprav k počtu nalezených duplicit kopírované věty hodnoceným systémem (tj. pokud systém najde větu bez úprav, měl by najít i větu po úpravě).

Tabulka 19 Míra platnosti hypotéz (zdroj: autor)

Hodnota kritéria	Závěr hypotézy
≥ 67	hypotéza potvrzena
≥ 33 a < 67	hypotézu nebylo možno potvrdit ani vyvrátit
<33	hypotéza vyvrácena

Tabulka 20 Zhodnocení hypotéz o úspěšnosti detekce (zdroj: autor)

Hypotéza	THE	TUR	EPH	GPL	Průměr
1	12%	40%	2%	56%	28%
2	14%	42%	6%	46%	27%
3	100%	100%	0%	0%	50%
4	100%	100%	0%	80%	70%
5	67%	100%	0%	4%	43%
6	0%	88%	na	0%	29%
7	0%	0%	0%	0%	0%
8	10%	50%	10%	30%	25%
9	0%	0%	0%	0%	0%
10	0%	40%	0%	70%	28%
11	20%	50%	0%	80%	38%

Na základě zhodnocení hypotéz v tabulce 20 a po porovnání jednotlivých systémů můžeme formulovat následující závěry o úspěšnosti detekce jednotlivých systémů:

Ani jeden testovaný systém neumí odhalit použité zdroje dokonale. Systémy Turnitin a GooglePlagiarism však dosahují výrazně lepších výsledků při odhalování zkopírovaných vět či odstavců (hypotézy 1 a 2). Systém MUNI – Theses.cz dosahuje horších výsledků pravděpodobně vlivem nastavení minimální hranice podobnosti 5 % obsahu dokumentu. V testu nejhůře dopadl systém Ephorus, který našel minimum dokumentů prokazujících plagiátorství (viz tabulka 16).

U systémů MUNI a Turnitin byly potvrzeny hypotézy 3 a 5. Pro úspěšnou detekci nevadí, pokud plagiátor přidá/odebere/parafrázuje slovo v kopírované větě. Naopak systémy Ephorus a GooglePlagiarism duplicitu v tomto případě nedetekují. V případě parafrázování celé věty (hypotéza 6) testované u cizojazyčných dokumentů dosahuje dobrých výsledků detekce pouze systém Turnitin.

Kromě systému Ephorus ostatním aplikacím nevadí, pokud je při kopírování z textu odstraněna diakritika (hypotéza 4).

Ani jedna aplikace nedokáže detekovat text přeložený z cizího jazyka, ani podpora překladů u systému Turnitin (v beta-verzi) neprokázala přínos při detekci. Hypotéza 7 nebyla potvrzena pro žádný z testovaných systémů.

Vzhledem k 5% hranici detekce u systému Theses.cz nebyla potvrzena hypotéza 8, že by tento systém byl nejlepší pro vyhledávání českých eVŠKP (úspěšnost 10 %). Pro tyto účely se lepšími ukázaly systémy Turnitin (50 %) a GooglePlagiarism (30 %).

Potvrdilo se, že systémy budou mít problémy s detekcí duplicit ze zdroje Anopress (hypotéza 9). Ve vyhledávání Open Access materiálů a EIZ si velmi dobře vedl především systém GooglePlagiarism využívající vyhledávač Google (úspěšnost 70 %), následován systémem Turnitin (úspěšnosti 40 %) – jedná se tedy o systémy využívající zahraniční indexy. Vzhledem k nízké obecné úspěšnosti aplikace Ephorus však nelze jednoznačně potvrdit hypotézu 10, že by všechny zahraniční aplikace dosahovaly u Open Access a webových zdrojů výsledků lepších.

Potvrdila se hypotéza 11, že autorova aplikace GooglePlagiarism využívající služeb webového vyhledávače dosahuje velmi dobrých výsledků u webových zdrojů (úspěšnost 80 % oproti druhému Turnitin s 50% úspěšností).

8.5 Závěr kapitoly

Systém Ephorus nelze, především díky nízké schopnosti detekce, doporučit pro kontrolu českých ani anglických textů.

Použití systémů Masarykovy univerzity Theses.cz a Odevzdej.cz se jeví jako vhodný kompromis mezi cenou a požadovanou úspěšností detekce. U těchto aplikací autor doporučil:

1) zrychlit vyhledávání,
2) odstranit omezení 5% hranice detekce,
3) zpřehlednit zobrazení nalezených duplicit.

Uvedená omezení byla vyřešena ve verzi Theses.cz zpřístupněné uživatelům na podzim 2014. Systém Theses.cz v nové verzi má mj. zrychlené vyhledávání, ale pro změnu chybně zobrazuje počet nalezených záznamů (viz oddíl 2.3.2 *Národní registr VŠKP – Theses.cz*).

Aplikace Turnitin dosahuje celkově velmi dobrých výsledků ve vyhledávání, má nejpropracovanější uživatelské rozhraní, proti nasazení aplikace však hovoří velmi vysoká cena za licenci.

Pro zlepšení detekce duplicit systémem GooglePlagiarism jsou na základě komparativní analýzy testovaných zdrojů a podle výsledků testu doporučeny následující úpravy aplikace:

a) implementovat vhodnější dělení textu do vět (hypotézy 1 a 2 – problém s odhalením věty/odstavce),
b) implementovat kontrolu kratších úseků než je celá věta (hypotézy 3 a 5 – problém s parafrází/úpravou jednoho slova),
c) zrychlit kontrolu vyhodnocování duplicit implementací ukládání nalezených dat do pomocného indexu, příp. paralelizací vyhledávání na více počítačů,
d) zlepšit zachování formátování zdrojového dokumentu při vyhodnocování duplicit pro lepší vizuální orientaci v textu.

9 Validátor VŠE

Při přípravě pokračování centralizovaného rozvojového projektu národního registru VŠKP pro rok 2011 autor disertační práce poukázal na potřebu zjednodušení zpřístupnění výsledků kontroly detekce na duplicity z externích systémů (viz kapitola 5), zjednodušení procesu vyhodnocování nalezených podezření na plagiátorství u eVŠKP a potřebu zavedení vyhledávání duplicit u dalších odborných dokumentů vznikajících na VŠE v Praze.

Tato kapitola formou případové studie popisuje předprojektovou přípravu, dílčí část projektu *Meziuniverzitní síť technických a metodických opatření na ochranu proti plagiátorství* realizovaného v roce 2011 a výslednou funkcionalitu popisované aplikace Validátor VŠE (http://validator.vse.cz) po více jak třech letech provozu. Případová studie tak může být inspirací na obdobný projekt pro ty vysoké školy, které výsledky kontroly na duplicity odpovědným osobám, oponentům kvalifikačních prací, doposud nezpřístupňují (viz 3.4.5 *Problematika plagiátorství*).

Cílem popisovaného projektu byla příprava uživatelského rozhraní mezi repozitáři VŠE v Praze a systémy Masarykovy univerzity – vybudování WWW aplikace Validátor VŠE, která by vyučujícím vhodnou formou zpřístupnila výsledky vyhledávání duplicitních pasáží (tj. podezření na plagiátorství) v plných textech eVŠKP a v dalších odborných publikacích vznikajících na VŠE v Praze. Výsledky kontroly jsou po vyhodnocení v externím systému aplikací staženy a zpřístupněny odpovědným osobám VŠE v Praze (vedoucí prací, oponenti, autoři textů apod.) prostřednictvím WWW rozhraní univerzity. Aplikace Validátor VŠE modulárně umožňuje zapojit další systémy na detekci duplicit, neboť samotné systémy MUNI mají svá omezení (jak vyplývá z kapitol 5 a 8).

9.1 Předprojektová příprava

Na začátku projektu bylo potřeba provést nejprve revizi analýzy potřeb školy (mezi podáním a schválením projektu uplyne doba 6 měsíců) a získat bližší zkušenosti o řešení problematiky plagiátorství na dalších vysokých školách formou dotazníkového šetření mezi účastníky semináře projektu Theses.cz.

V rámci návrhu projektu byly autorem disertační práce řešeny mj. importy a exporty metadat, plných textů, přístupová oprávnění, finanční, organizační a personální zajištění provozu aplikace a provázanost na externí systémy VŠE v Praze a MUNI, jak je popsáno níže.

9.1.1 Analýza výchozího stavu

První centralizovaný rozvojový projekt vysokých škol na vybudování národního registru VŠKP Theses.cz a kontrolu plagiátorství byl přijat MŠMT ČR pro rok 2008. V rámci projektu VŠE v Praze implementovala WWW aplikaci Databáze kvalifikačních prací VŠE pro odevzdávání a správu elektronických kvalifikačních prací, s exportem metadat do knihovního katalogu Aleph a exportem metadat a plných textů do Theses.cz.

V dalších letech bylo podáno několik navazujících centralizovaných rozvojových projektů, které rozvíjely původní projekt z roku 2008. Tyto navazující projekty řešily repozitáře seminárních prací (web http://www.odevzdej.cz), dalších vědeckých prací (web http://www.repozitar.cz) a související vyhledávání duplicit v těchto pracích.

Na VŠE v Praze bylo odevzdávání plných textů eVŠKP převedeno z Databáze kvalifikačních prací VŠE do nového školního informačního systému ISIS a v tomto systému implementováno zpřístupnění výsledků kontroly seminárních prací vyučujícím. Z ISIS jsou metadata eVŠKP exportována do knihovního katalogu knihovny Aleph a do původní Databáze kvalifikačních prací VŠE.

Výsledky kontroly eVŠKP ze systému Theses.cz byly dostupné pouze autorizovanému správci systému za VŠE v Praze Janu Machovi. Nebylo proto možné pravidelně a efektivně vyhodnocovat jednotlivá podezření na plagiátorství u všech eVŠKP obhajovaných na univerzitě.

Po roce 2009 se zúčastněné vysokoškolské knihovny zaměřily na podporu otevřeného přístupu k vědeckým informacím. Vzhledem k tomuto novému trendu podpory Open Access publikování se ukázala potřeba vybudování univerzitního repozitáře vědeckých textů (aplikace Repozitář VŠE) a vhodnost povinného vyhledávání podobných textů u odborných vědeckých textů publikovaných autory z VŠE v Praze, příp. i u libovolných dalších textů (např. připravované publikace, časopisecké články, průběžná kontrola odevzdávaných prací studenty).

VŠE v Praze započala v roce 2010 s virtualizací provozovaných systémů a aplikací na architektuře IBM BladeCenter, což ovlivnilo následné rozhodnutí o potřebné hardwarové architektuře pro projekt Validátor VŠE.

Na základě analýzy výchozího stavu bylo rozhodnuto podat centralizovaný rozvojový projekt pro rok 2011, jehož součástí bylo vybudování aplikace Validátor VŠE zajišťující zpřístupnění výsledků kontroly eVŠKP a dalších prací odpovědným osobám.

9.1.2 Odpovědnost za realizaci projektu, organizační zajištění

Vzhledem k nutnosti dohadování formálních stránek projektu s hlavním řešitelem MUNI Michalem Brandejsem byl na VŠE v Praze hlavním řešitelem projektu zvolen ředitel výpočetního centra Karel Nenadál. Další členové projektového týmu byly:

Projektový manažer:	Jan Mach
Odpovědná osoba za rozpočet projektu:	Dana Václavíková

Kromě výše uvedených osob na projektu dle potřeby spolupracovali další osoby z Centra informačních a knihovnických služeb (Repozitář VŠE) a Výpočetního centra (systém ISIS, HW zajištění).

Pro osoby pracující nad rámec svých běžných pracovních povinností bylo v projektu požadováno financování osobního ohodnocení. Kromě mzdových nákladů obsahovala kalkulace projektu Odvody pojistného na veřejné zdravotní pojištění a pojistného na sociální zabezpečení a příspěvku na státní politiku zaměstnanosti a příděly do sociálního fondu v odpovídající procentuální výši[41].

9.1.3 Koncepce navrhovaného systému

Na základě analýzy výchozího stavu a potřeb autor disertační práce připravil koncepci navrhovaného systému – funkce, kritéria úspěšnosti, alternativní scénáře realizace a kritéria

[41] Procento odvodů bylo změněno mezi podáním projektu v roce 2010 a realizací projektu v roce 2011, proto bylo nutné v průběhu projektu požádat MŠMT ČR o úpravu rozložení finančních prostředků do jednotlivých kapitol rozpočtu.

pro výběr. Pro provoz systému se ukázala potřeba posílení HW infrastruktury na bázi IBM BladeCenter o nový blade server.

Hlavní a vedlejší funkce systému

Hlavní funkce systému:

- automatický export metadat pro potřeby vyhledávání duplicit u eVŠKP,
- automatický import výsledků vyhledávání duplicit u eVŠKP,
- zpřístupnění výsledků vyhledávání duplicit eVŠKP odpovědným osobám (vyučujícím, oponentům, příp. vedoucím a sekretářkám kateder).

Volitelné funkce systému:

- napojení na repozitář vědeckých textů VŠE v Praze, vyhledávání duplicit u fulltextových dokumentů v Repozitáři VŠE,
- notifikace uživatelů e-mailem o ukončení kontroly dokumentů a dostupnosti výsledků ve Validátoru VŠE,
- podpora vyhledávání projevů plagiátorství u libovolných dalších textů nahraných zaměstnancem nebo doktorandem.

9.1.4 Alternativní scénáře realizace

Při přípravě projektu byly autorem vypracovány tři alternativní scénáře kontroly odborných prací na projevy plagiátorství a stanoveny jejich hlavní klady a zápory:

1) Zpřístupnění rozhraní systémů MUNI Theses.cz (pro kontrolu eVŠKP) a Odevzdej.cz (pro kontrolu libovolných dalších dokumentů) přímo vyučujícím.
    - výhody: malá náročnost na implementaci
    - nevýhody: použity dvě odlišné externí aplikace, s uživatelsky málo přívětivým uživatelským rozhraním, s velmi omezenou možností přizpůsobení, komplikovaná správa přístupových práv, omezení pouze na určitý typ dokumentů (eVŠKP nebo seminární práce, komplikovaně řešitelná integrace dat z Repozitáře VŠE)

2) Integrace nové funkcionality do školního informačního systému ISIS, kde již bylo obdobným způsobem řešeno vyhledávání duplicit u seminárních prací.
    - výhody: systém známý cílové uživatelské skupině, ISIS již implementoval kontrolu seminárních prací, odpadá nutnost exportu dat eVŠKP, provoz na serverech VŠE v Praze
    - nevýhody: finančně velmi nákladná implementace i provoz v následujících letech, komplikovaně řešitelná integrace dat z Repozitáře VŠE, omezenější nebo finančně nákladná možnost přizpůsobení (úpravy aplikace mohou provádět pouze externí vývojáři)
3) Zhotovení samostatné aplikace s možností propojení na ISIS.
    - výhody: vysoká flexibilita systému, nízké provozní výdaje, možnost následných úprav aplikace interně na VŠE v Praze, provoz aplikace na serverech univerzity
    - nevýhody: samostatná aplikace (i když s možností implementace jednotného uživatelského rozhraní používaného univerzitou)

9.1.5 Kritéria výběru scénáře

Na základě znalosti výchozího stavu a výsledků analýzy potřeb byla v rámci projektového týmu stanovena následující kritéria pro výběr scénáře fungování aplikace:

1) podpora požadovaných hlavních a vedlejších funkcí systému,
2) uživatelská přívětivost systému včetně možnosti přizpůsobení designu,
3) náklady na vývoj a následnou údržbu systému,
4) jednoduchost úprav v následných letech.

Z výše uvedených důvodů byl již v předprojektové přípravě zamítnut první ze scénářů, tj. zpřístupnění systémů MUNI přímo akademické obci VŠE v Praze.

9.1.6 Studie proveditelnosti

Před podáním projektu projektový tým při zpracování studie proveditelnosti oslovil se žádostí o upřesnění/odsouhlasení možností systému a nacenění požadované funkcionality tyto subjekty:

1) správce ISIS ve Výpočetním centru VŠE v Praze,
2) programátorský tým Mendelovy univerzity, který spravuje školní informační systém ISIS,
3) vývojáře Databáze kvalifikačních prací VŠE.

Studie proveditelnosti přispěla mj. ke stanovení funkcionalit, časového harmonogramu a rozpočtu projektu.

9.1.7 Časový harmonogram

Projekt byl naplánován, i vzhledem ke zvolenému dotačními programu, na období leden – prosinec 2011 s tím, že uvedení aplikace Validátor VŠE do plného provozu bude možné až v roce následujícím. V rámci celého projektu na rok 2011 bylo kromě Validátoru VŠE řešeno vybudování lokálního Repozitáře VŠE pro vědecké texty a jeho napojení na připravovaný server MUNI Repozitar.cz.

Tabulka 21 obsahuje navržený harmonogram prací na Validátoru VŠE, vypracovaný s ohledem na zvolený dotační program, náročnost jednotlivých etap, čerpání dovolených na vysokých školách převážně v letních měsících a na požadavek VŠE v Praze na interní vyúčtování projektu do konce listopadu 2011.

Tabulka 21 Časový harmonogram prací na Validátoru VŠE (zdroj: autor)

Příprava projektu	leden – březen 2011
Výběr dodavatele, zadávací dokumentace, smlouvy	březen – květen 2011
Programování aplikace, posílení HW infrastruktury	květen – září 2011
Ověřovací provoz, příprava metodických opatření na VŠE	září – říjen 2011
Předání a vyúčtování projektu	říjen – listopad 2011
Schválení organizačních opatření, předání aplikace do provozu, reportování projektu	listopad - prosinec 2011

V průběhu samotného projektu se ukázalo vhodným zpracovat Ganttův diagram (viz Příloha XIII) s detailnějším členěním jednotlivých etap projektu, příp. s jejich návaznostmi, náročností a stupněm plnění.

9.1.8 Rozpočet projektu

Tabulka 22 obsahuje strukturování rozpočtu projektu, dané zvoleným dotačním programem MŠMT ČR. V tabulce je kurzívou uvedeno, na které činnosti z popisované dílčí části projektu Validátor VŠE byla dotace požadována.

Tabulka 22 Strukturování rozpočtu projektu (zdroj: autor)

1. Kapitálové finanční prostředky	
1.1	Dlouhodobý nehmotný majetek (SW, licence) *- Validátor VŠE (modul ISIS nebo samostatná aplikace)*
1.2	Samostatné věci movité (stroje, zařízení) *- posílení HW infrastruktury*
1.3	Stavební úpravy
2. Běžné finanční prostředky	
Osobní náklady:	
2.1	Mzdy (včetně pohyblivých složek) *- odměny řešitelům (analýza, příprava zadání a testování aplikace, tvorba návodů a organizačních opatření pro odevzdávání a zpřístupnění zaměstnaneckých děl, analýza, příprava zadání a testování aplikace, projektový manažer)*
2.2	Odměny dle dohod o pracích konaných mimo pracovní poměr
2.3	Odvody pojistného na veřejné zdravotní pojištění a pojistného na sociální zabezpečení a příspěvku na státní politiku zaměstnanosti a příděly do sociálního fondu *- vypočítáno jako podíl z mezd*
Ostatní:	
2.4	Materiální náklady (včetně drobného majetku)
2.5	Služby a náklady nevýrobní
2.6	Cestovní náhrady
2.7	Stipendia

V projektu byla u jednotlivých položek uvedena požadovaná výše dotace a detailněji zdůvodněna např. uvedením konkrétní cenové nabídky. Celkovou výši dotace na celý projekt VŠE v Praze, zahrnující mj. prostředky na Validátor VŠE, zobrazuje Tabulka 23 *Přidělené prostředky na projekt*.

Tabulka 23 Přidělené prostředky na projekt (zdroj: autor)

Neinvestiční prostředky:	519 tis. Kč
Investiční prostředky:	820 tis. Kč
Celkem:	**1 339 tis. Kč**

9.2 Projektový úkol Meziuniverzitní síť technických a metodických opatření na ochranu proti plagiátorství

Projektový tým VŠE v Praze vypracoval žádost o dotaci v předepsané struktuře podle formuláře MŠMT ČR. Návrh projektu, jehož součástí bylo i popisované řešení Validátoru VŠE, byl přijat jako centralizovaný rozvojový projekt ROZV/C39/2011 *Meziuniverzitní síť technických a metodických opatření na ochranu proti plagiátorství*. Dotační program a jednotlivé části projektu související s Validátorem VŠE jsou popsány níže, příp. již byly nastíněny v rámci předprojektové přípravy (analýza výchozího stavu, požadavky na cílové chování systému, časový harmonogram, rozpočet projektu).

9.2.1 Zvolený dotační program

Řešitelé MUNI ve spolupráci se spoluřešiteli z jednotlivých zapojených vysokých škol zvolili jako zdroj financování centralizované rozvojové programy MŠMT na rok 2011. Rozvojové programy vyhlašuje MŠMT ČR jednou ročně, s uzávěrkou přihlášek obvykle v říjnu před rokem, pro který je program vyhlášen. „Cílem rozvojových programů je přispět k naplňování jednotlivých priorit stanovených v Dlouhodobém záměru ministerstva a jeho Aktualizaci.“ (110 str. 1)

Rozvojové programy se dělí na dva okruhy – na institucionální rozvojové programy (určené pro jednotlivé veřejné vysoké školy; v roce 2011 nazývané decentralizované rozvojové projekty) a na centralizované rozvojové programy, které byly využity pro financování popisovaného projektu.

„V rámci centralizovaných rozvojových programů mohou být veřejným vysokým školám poskytovány dotace s přihlédnutím k výsledku hodnocení předložených projektových žádostí, které provede Rada programů. Výsledkem hodnocení může být doporučení úprav projektu a krácení požadovaných finančních prostředků.

Předkládané projekty veřejných vysokých škol budou založeny na analýze jejich činností a výsledků za uplynulé období. V roce 2013 budou prostřednictvím centralizovaných rozvojových programů podpořeny aktivity uskutečňované dvěma formami:

1. Projekty konsorcií vysokých škol, kdy centralizovaný projekt musí podat společně alespoň dvě vysoké školy. V programech 2 a 3 mohou vysoké školy předkládat projekty i samostatně.
2. Projekty vysokých škol se sídlem na území hlavního města Prahy, překládané buď samostatně, nebo jako projekty konsorcií (program č. 3).“ (110 str. 2)

Pro projekt byl zvolen dotační program 6 pro rok 2011 s názvem *Program na podporu dalších aktivit vysokých škol*, podprogram *e) podprogram na podporu kontroly a ochrany proti plagiátorství*. MUNI jakožto programátor systémů Theses.cz a Odevzdej.cz se stala stejně jako i v minulých letech koordinátorem konsorcia vysokých škol.

Kromě koordinátora – Masarykovy univerzity – se na projektu podílely:

1. Česká zemědělská univerzita
2. Janáčkova akademie múzických umění v Brně
3. Jihočeská univerzita v Českých Budějovicích
4. Ostravská univerzita v Ostravě
5. Slezská univerzita v Opavě
6. Technická univerzita v Liberci
7. Univerzita Jana Evangelisty Purkyně
8. Univerzita Karlova v Praze
9. Univerzita Palackého v Olomouci
10. Vysoká škola báňská-Technická univerzita
11. Vysoká škola ekonomická v Praze
12. Vysoká škola polytechnická Jihlava
13. Vysoká škola technická a ekonomická
14. Západočeská univerzita v Plzni

Centralizované rozvojové projekty se ukázaly jako vhodný dotační program již při budování národního registru VŠKP Theses.cz v roce 2008 a projektů souvisejících v letech 2009 a 2010. Pravděpodobně z tohoto důvodu byl ministerstvem vypsaný dotační podprogram pro rok 2011 přímo zaměřen na problematiku plagiátorství.

9.2.2 Zadání, cíle projektu

Abstrakt celého podávaného projektu zmiňuje hlavní cíle a řešené oblasti na jednotlivých školách:

„Cílem projektu Meziuniverzitní síť technických a metodických opatření na ochranu proti plagiátorství patnácti veřejných vysokých škol je vytvořit technická a metodická opatření na kontrolu a ochranu proti plagiátorství. Jde o vytvoření nástrojů i norem, které vysoké školy budou moci využívat ve všech fázích ochrany proti plagiátorství, tj. pro předcházení plagiátorství, pro vyhledávání (nalezení) plagiátů a pro řešení pozitivních nálezů (případů). Tyto nástroje a metodiky vzniknou na základě co nejširší spolupráce a sdílení zkušeností se všemi zapojenými školami. Z analýzy potřeb do projektu zapojených škol vyplývají následující cíle:

1. vytvoření metodik, pravidel a postupů pro sběr prací od autorů,
2. vytvoření resp. zlepšení technických podmínek pro sběr prací a kontrolu textů,
3. tvorba pravidel upřesňujících postup při pozitivním nálezu podobných textů,
4. vytvoření nástrojů pro statistické zhodnocení úspěšnosti procesu.“ (111)

Dílčí část projektu za VŠE v Praze kromě vypracování popisované aplikace obsahovala tyto cíle:

1. analýza potřeb školy (resp. revize analýzy potřeb školy z III. čtvrtletí 2010),
2. pracovní setkání spoluřešitelů k řešení projektu, klíčových kroků projektu a řešení autorskoprávních otázek s právními odborníky aj.,
3. seznámení se s Koncepcí projektu, příp. připomínky ke Koncepci projektu,
4. výměna zkušeností k procesním systémům odhalování plagiátů se zástupci z vysokých škol prostřednictvím semináře,
5. dvoustranné smlouvy o spolupráci,
6. postupy, metodiky a pravidla pro sběr prací od autorů a pro řešení postupů při pozitivním nálezu podobných textů (dle analýzy),
7. zhodnocení úspěšnosti kontroly plagiátorství v závěrečných pracích,
8. pořízení a posílení hardware pro lokální systém,
9. příprava lokálního úložiště školy pro sběr zaměstnaneckých a doktorských děl a napojení lokálního úložiště školy na Repozitar.cz,

10. vložení prací školy do Repozitar.cz,
11. vyhodnocení projektu za rok 2011.

9.2.3 Hypotéza výsledného chování projektovaného IS

Na základě požadavků na výsledné chování projektovaného informačního systému Validátor VŠE byl z možných alternativních hypotéz o chování systému (viz 9.1.4 *Alternativní scénáře realizace*) vybrán scénář č. 3 – vývoj samostatné aplikace. Vybraná hypotéza chování systému byla dále rozpracována. Důvodem tohoto výběru byly velmi vysoké finanční náklady na doprogramování odpovídajícího modulu do školního informačního systému ISIS a následné každoroční servisní poplatky ve výši 20-25 % z pořizovací ceny u scénáře č. 2; scénář č. 1 byl zamítnut již v předprojektové přípravě. Výhodou samostatné aplikace podle scénáře č. 3, kromě stránky finanční, je větší míra poskytované funkcionality a jednodušší možnosti následných úprav (přizpůsobení) např. z důvodu možných budoucích změn externích aplikací MUNI[42].

Požadovaná funkcionalita:

- automatizovaný import metadat a příp. plných textů z externích systémů (ISIS/Databáze kvalifikačních prací VŠE, Repozitář VŠE, WWW rozhraní pro uživatele),
- uživatelské rozhraní pro vkládání vlastních fulltextových souborů uživateli,
- export metadat a plných textů do externích systémů za účelem kontroly duplicit (Theses.cz, Odevzdej.cz, příp. modulárně rozšiřitelné pro případ potřeby),
- import výsledků provedených analýz z externích systémů,
- notifikace relevantních uživatelů e-mailem o dostupnosti zpracované analýzy,
- možnost uživatelsky třídit jednotlivé záznamy do složek,

[42] Příkladem změny externí aplikace Theses.cz je naprogramování funkcionality výpočtu souhrnné podobnosti za více dokumentů najednou. API pro napojení na Validátor VŠE bylo uvolněno na podzim 2014, na začátku roku 2015 byla tato funkcionalita doprogramována do Validátoru VŠE. Před provedenou změnou umožňoval systém Theses.cz zjistit pouze míru podobnosti mezi dvěma dokumenty, což bylo i předmětem kritiky v analýze antiplagiátorských systémů viz kapitola 8.

- filtrování, třídění a vyhledávání jednotlivých záznamů,
- přehledná informace o průběhu či výsledku analýzy dokumentů relevantním uživatelům.

Požadavky na přenositelnost:

- uživatelské rozhraní WWW/HTML/CSS,
- programovací jazyk PHP,
- unixový operační systém Solaris,
- import dat protokolem OAI-PMH, použití webových služeb,
- standardy Dublic Core, EVSKP-MS, příp. další specifické formáty MUNI,
- modularita exportních a importních funkcí pro umožnění rozšíření aplikace o další vstupně/výstupní systémy.

Databázové systémy:

- MySQL (Databáze VŠKP, Validátor VŠE),
- Oracle (ISIS).

Integrace produktů třetích stran:

- ISIS (import dat a číselníků o VŠKP, příp. export výsledků kontroly do ISIS),
- Databáze kvalifikačních prací VŠE (import metadat o obhajovaných VŠKP),
- Repozitář VŠKP (import metadat a plných textů evidovaných prací z Repozitáře VŠKP),
- Theses.cz (kontrola VŠKP na projevy plagiátorství),
- Odevzdej.cz (kontrola ostatních prací na projevy plagiátorství).

9.2.4 Zúčastněné subjekty

V rámci projektu byly zapojeny tyto subjekty:

- Zadavatel projektu: Vysoká škola ekonomická v Praze
- Poskytovatel dotace: MŠMT ČR
- Hlavní řešitel VŠE: Karel Nenadál, Výpočetní centrum VŠE v Praze
- Projektant, manažer projektu, správce aplikace: Jan Mach, CIKS

- Kooperant projektanta: Jan Říha, Výpočetní centrum VŠE v Praze
- Realizátor projektu: DB GROUP s.r.o
- Provozovatel systému: Výpočetní centrum VŠE v Praze
- Spolupracující instituce: Mendelova univerzita v Brně (ISIS), Masarykova univerzita (Theses.cz, Odevzdej.cz, Repozitar.cz)

Na VŠE v Praze byly identifikovány následující role uživatelů systému:

- doktorandi (kontrola vlastní publikační činnosti)
- vyučující, vědečtí pracovníci (kontrola vlastní publikační činnosti a vedených VŠKP)
- oponenti (kontrola oponovaných VŠKP, oponent nemusí být zaměstnancem VŠE v Praze)
- vedoucí katedry, sekretářka katedry (zastupitelnost výše uvedených uživatelských skupin)
- správci systému Validátor VŠE (dohled nad provozem, řešení uživatelských a technických problémů)
- správci externích systémů (import/export dat ze systémů Repozitář VŠE, ISIS a Databáze VŠKP)
- administrátor Validátoru VŠE (údržba systému, konfigurace napojení na vstupně/výstupní systémy, řešení uživatelských problémů aj.)
- studenti (nejsou přímí uživatelé systému, iniciují workflow kontroly eVŠKP jejím vložením do systému ISIS)

9.2.5 Indikátory úspěšnosti projektu

V rámci projektu byly stanoveny následující indikátory úspěšnosti:

1) provedený import číselníků ze systému ISIS (uživatelé systému aj.)
2) probíhající import metadat a plných textů
    a) eVŠKP
    b) vědecké texty - Repozitář VŠE
    c) uživatelské zadávání vlastních dokumentů
3) zpracování metodických pokynů a prezentace výsledku projektu cílené uživatelské skupině
4) spuštění aplikace do ostrého provozu

Za úspěšné splnění cílů projektu bude považováno splnění indikátorů 1 - 3, pro všechny typy fulltextových souborů a) – c). Vzhledem k omezení projektu pouze na rok 2011 nelze předpokládat, že bude možné převedení projektu do ostrého provozu (indikátor 4) již na konci roku 2011 z důvodu nutnosti přijetí potřebných organizačně provozních pravidel na škole.

Vypracovaný Ganttův diagram (viz Příloha XIII) definuje dílčí mezníky harmonogramu důležité pro pokračování následujících fází projektu.

9.2.6 Organizačně funkční řešení systému

S vybraným realizátorem projektu a dalšími zúčastněnými stranami VŠE v Praze byl zpracován síťový graf zachycující jednotlivé vstupní/výstupní systémy a tok plných textů a metadat (viz Příloha XIV) a organizačně funkční schéma projektu (resp. sada vývojových diagramů, ukázka jednoho z diagramů viz Příloha XV). Součástí přílohy smlouvy byl návrh jednotlivých hlavních tříd aplikace připravený autorem disertační práce.

9.2.7 Typy dokumentů

Aplikace kontroluje následující typy dokumentů prostřednictvím uvedených systémů MUNI na vyhledávání duplicit (jednotlivá použitá rozhraní upřesněna níže):

1) vysokoškolské kvalifikační práce – vstupní rozhraní: ISIS nebo Databáze kvalifikačních prací VŠE, kontrola na plagiátorství: systém Theses.cz
2) vědecké texty – vstupní rozhraní: Repozitář VŠE (OAI-PMH server), kontrola na plagiátorství: systém Odevzdej.cz nebo Repozitar.cz
3) uživatelské dokumenty – vstupní rozhraní: WWW stránky aplikace, kontrola na plagiátorství: systém Odevzdej.cz

9.2.8 Vstupní a výstupní rozhraní systému

Vstupní rozhraní definuje systémy, do kterých jsou ručně vkládána nebo ze kterých jsou automatizovaně sklízena metadata a plné texty dokumentů. Aplikace umožňuje dodatečnou definici dalších, níže neuvedených vstupních rozhraní díky otevřenému rozhraní aplikace.

WWW stránky aplikace

WWW stránka Validátor VŠE umožňuje, po autentifikaci a autorizaci uživatele systémem Shibboleth, vložit metadata a plný text práce ke kontrole.

ISIS

Studijní informační systém ISIS generuje na předem dané URL adrese soubory obsahující XML metadatové záznamy ve formátu EVSKP-MS (1) o eVŠKP před obhajobou a následně změněná metadata po obhajobě.

OAI-PMH

Protokol OAI-PMH (viz podkapitola 4.3 *Implementace OAI-PMH serveru na VŠE v Praze*) zpřístupňuje metadatové záznamy, které obsahují URL odkaz na plný text a identifikaci autorů/oprávněných osob. Aplikace Validátor VŠE umožňuje v intervalech definovaných správcem automaticky stahovat přírůstky – metadatové záznamy – z OAI-PMH serveru a importovat tak záznamy pro kontrolu duplicit.

Repozitář VŠE

Repozitář VŠE je připravovanou konkrétní implementací OAI-PMH serveru s metadaty a plnými texty vědeckých prací VŠE.[43]

Rozhraní pro kontrolu duplicit

Kontrola na výskyt duplicit je realizována v rámci projektu prostřednictvím těchto systémů:

- aplikace Theses.cz
    - kontrola eVŠKP; export metadat a plných textů již probíhala z Databáze kvalifikačních prací VŠE a nebyla proto předmětem tohoto projektu
- aplikace Repozitar.cz
    - vědecké práce, automaticky přebírané z Repozitáře VŠE

[43] V roce 2014 byl projektový záměr ukládání publikační činnosti na VŠE v Praze změněn, jako primární platforma pro ukládání byl zvolen knihovní systém Aleph. Vzhledem ke vhodně zvolené modulární funkcionalitě Validátoru VŠE tato změna nebude mít vliv na činnost aplikace ani na uživatelskou zkušenost.

- aplikace Odevzdej.cz
    - primárně určeno pro seminární práce, v projektu použito pro kontrolu libovolných prací zadaných uživateli ve Validátoru VŠE

Na základě ověřovacího provozu může být v budoucnu rozhodnuto o rozšíření aplikace Validátor VŠE o další externí rozhraní na kontrolu projevů plagiátorství, např. prostřednictvím aplikace iPlagiarism[44].

Vstupně – výstupní rozhraní a toky dat (metadat a plných textů) jsou schematicky znázorněny na síťovém diagramu (viz Příloha XIV), jednotlivé procesy byly detailněji popsány v zadání aplikace v komentářích nebo vývojovým diagramem pro konkrétní procesy.

9.3 Výsledná aplikace Validátor VŠE

Na základě analýzy a zadání aplikace vypracované Janem Machem byla externí společností naprogramována aplikace Validátor VŠE (112). Aplikace po svém spuštění slouží vyučujícím a oponentům VŠE v Praze ke zpřístupnění výsledků kontroly detekce duplicit u eVŠKP a vlastních textů nahraných do aplikace. Kontrola probíhá prostřednictvím systémů MUNI Theses.cz (eVŠKP) a Odevzdej.cz (další texty). Vyučující jsou o výsledcích kontroly informováni e-mailem, s volitelnou minimální mírou nalezené podobnosti, při které je e-mail zasílán.

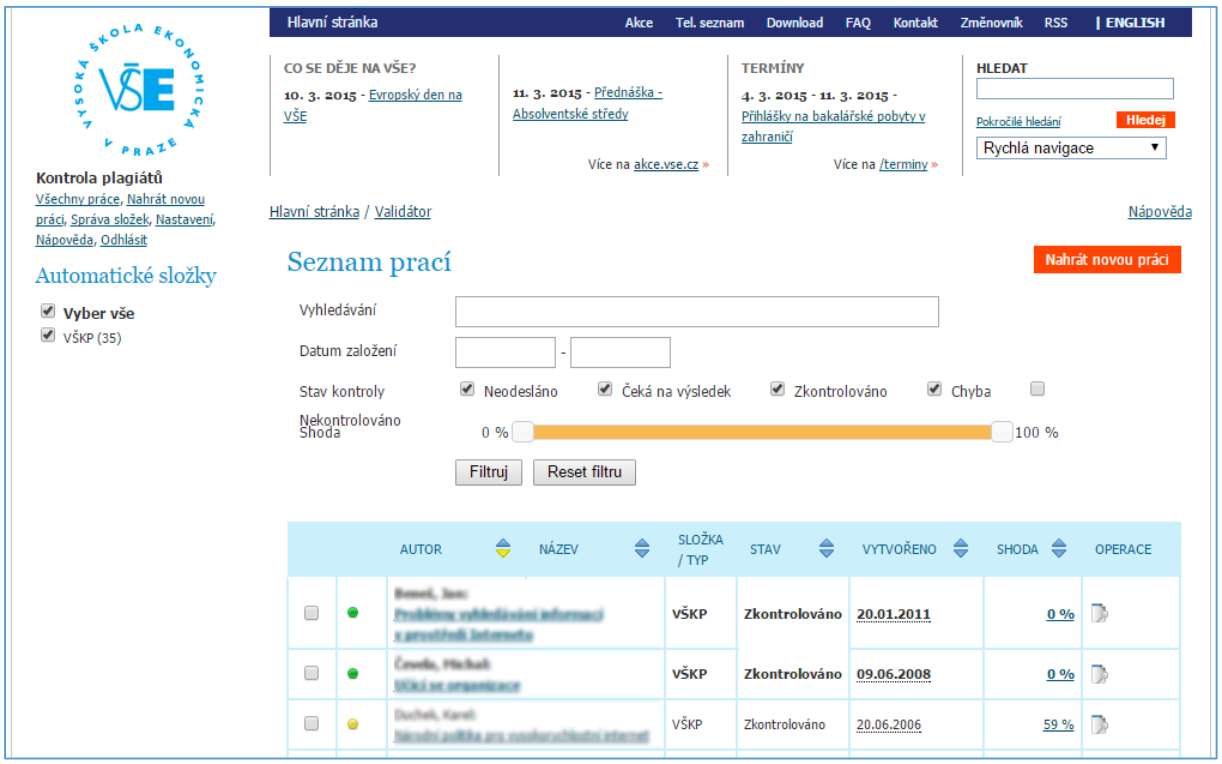
V průběhu vývoje a následného provozu aplikace byla autorem disertační práce navrhnuta řada expertních pravidel pro vyloučení chyby 1. druhu při detekci plagiátorství (falešně pozitivní duplicity, např. nalezena podobnost mezi odevzdanou diplomovou prací a seminární prací zpracovávanou shodným studentem v diplomovém semináři). Pravidla byla integrována v aplikaci Validátor VŠE. Po filtraci nalezených duplicit podle pravidel došlo k významnému snížení falešných hlášení o podezření plagiátorství ve Validátoru VŠE oproti výsledkům samotného systému Theses.cz.

Poslední realizovaná úprava Validátoru VŠE proběhla na přelomu let 2014/2015, kdy byly do aplikace zapracovány úpravy provedené na podzim 2014 v aplikačním rozhraní Theses.cz. Jedná se především o zobrazení souhrnného protokolu se všemi nalezenými podobnostmi,

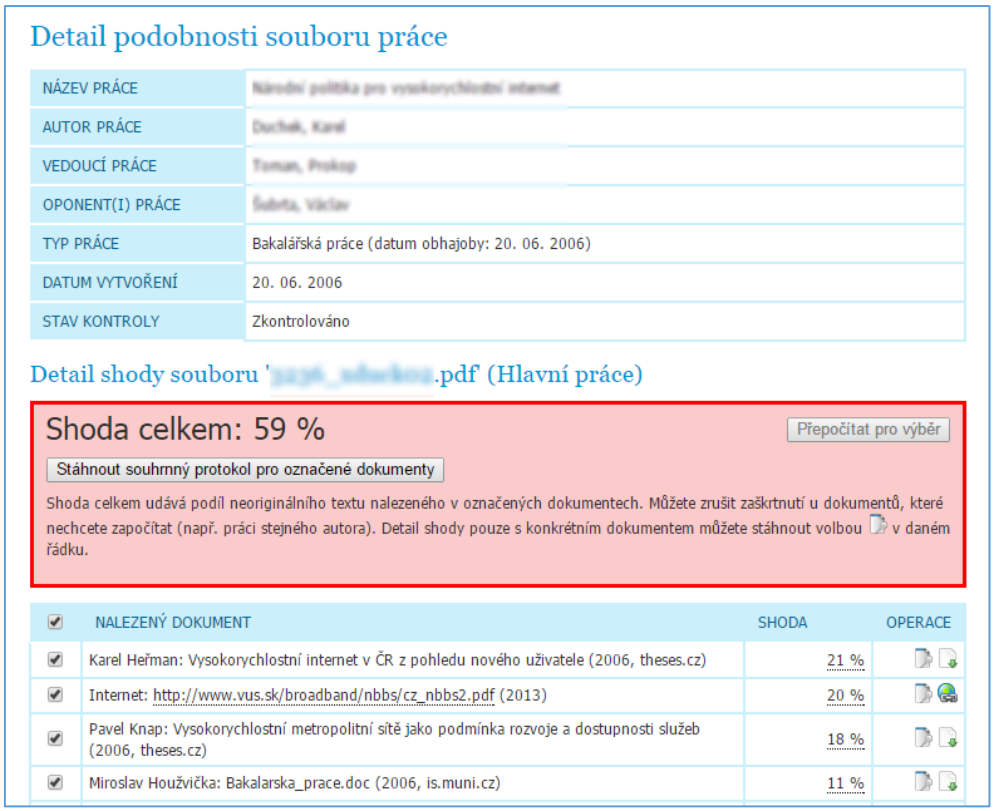
[44] VŠE v Praze zakoupila v roce 2015 službu kontroly preprintů časopiseckých článků v aplikaci iPlagiarism. Správcem aplikace za VŠE v Praze je Ing. Jan Mach, pověřeným i vyhodnocováním nalezených duplicit.

výpočet celkového podílu všech nalezených shod v textu analyzované práce a možnost výpočtu souhrnné podobnosti pouze za uživatelem vybrané dokumenty. Výpočet souhrnné podobnosti umožňuje uživatelům získat přehled o celkovém podílu shodných textů v celé analyzované práci i při velkém množství nalezených podobných dokumentů, bez nutnosti zvlášť kontrolovat protokol pro každý z nalezených dokumentů.

Aplikace Validátor VŠE je dostupná akademické obci na VŠE v Praze na adrese http://validator.vse.cz (viz Obrázek 27 a Obrázek 28, kvůli ochraně byly osobní údaje rozostřeny).



Obrázek 27 Validátor VŠE - přehled prací (112)



Obrázek 28 Validátor VŠE - detail podobnosti (112)

9.4 Závěr kapitoly

Výše uvedená případová studie na konkrétním příkladu popisuje projektovou přípravu a realizaci webové aplikace, zpřístupňující vhodnou formou nalezené duplicity v odevzdávaných vysokoškolských kvalifikačních a jiných pracích. Může tak posloužit jako inspirace vysokým školám, které výsledky kontroly eVŠKP z Theses.cz odpovědným osobám doposud nezpřístupňují (viz průzkum v oddílu 3.4.5 *Problematika plagiátorství*).

Po dobu provozu aplikace bylo na VŠE v Praze díky ní zachyceno řádově několik desítek eVŠKP obsahujících významné duplicity s externími dokumenty, vyhodnocené jako podezření na plagiátorství.

Vzhledem k tomu, že právní úprava neumožňuje odebrání titulu zpětně (viz oddíl 2.1.1 *Definice vysokoškolských kvalifikačních prací*), aplikace Validátor VŠE je významným nástrojem pro včasnou detekci pokusů o podvod při získání akademického titulu. Neméně významný je i preventivní účinek nasazení aplikace, kdy studenti vědí, že případné plagiátorství bude jednoduše odhaleno díky automatizaci detekce duplicit.

10 Závěr a přínosy disertační práce

Disertační práce poskytuje ověřená řešení problémů souvisejících se správou, zpřístupňováním a vyhledáváním eVŠKP v ČR. Konkrétně je řešena problematika mapování metadatových prvků, komunikace a vyhledávání záznamů eVŠKP, metrik pro měření užití záznamů eVŠKP v otevřených repozitářích a problematika plagiátorství u eVŠKP.

Mezi hlavní přínosy disertační práce patří:

1) Zpracovaný průzkum vybraných repozitářů eVŠKP v ČR a v zahraničí, analýza stavu zpřístupňování eVŠKP v ČR. Do dotazníkového šetření v roce 2014 byly poprvé zapojeny všechny veřejné vysoké školy.
2) Doporučení mapování metadatových prvků standardu EVSKP-MS do jiných metadatových sad, doporučení vhodné praxe komunikace metadat protokolem OAI-PMH a autorskoprávní analýza volného zpřístupnění eVŠKP na Internetu.
3) Analýza metrik užití eVŠKP v otevřených repozitářích, doporučení pro agregaci a zpracování statistických dat užití eVŠKP.
4) Řešení problematiky vyhledávání eVŠKP na příkladu modelové aplikace vyhledávacího serveru s fasetovým vyhledáváním a výběru discovery služby podporující vyhledávání záznamů eVŠKP.
5) Analýza antiplagiátorských řešení, doporučení zlepšení služeb a prezentace dobré praxe zpřístupňování výsledků kontroly na příkladu implementace aplikace Validátor VŠE.

Autor na začátku práce podrobně představuje a kriticky hodnotí výchozí stav sběru, zpřístupňování a vyhledávání vysokoškolských kvalifikační prací v ČR. Na základě komparační analýzy významných zahraničních a českých digitálních repozitářů VŠKP analyzoval kritickou funkcionalitu a trendy ve zpřístupňování eVŠKP. Význam průzkumů realizovaných za činnosti Komise eVŠKP pro praxi dokládá jejich užití např. v analyzovaných kvalifikačních pracích. Autor navázal na průzkumy Komise eVŠKP výzkumem stavu zpřístupňování VŠKP v roce 2014, kterého se poprvé v historii zúčastnily všechny veřejné vysoké školy. Výsledky průzkumu byly publikovány v recenzovaném časopise *ProInflow 2/2014* a jsou popsány ve 3. kapitole disertační práce.

Metadatový standard EVSKP-MS, příp. odvozený proprietární formát Theses.cz, slouží pro komunikaci metadat eVŠKP v ČR. V kapitole 4. autor formuloval doporučení pro mapování prvků metadatového standardu EVSKP-MS do dalších významných metadatových souborů. Mapování prvků EVSKP-MS je důležité např. při komunikaci metadat prostřednictvím protokolu OAI-PMH, jako je např. import metadat disertačních prací VŠE v Praze do repozitáře šedé literatury NUŠL nebo evropského repozitáře DART-Europe E-theses Portal.

Kapitola 5 kriticky analyzuje možnosti hodnocení užití eVŠKP v online prostředí za využití klasických citačních metrik a metrik založených na odezvě práce v sociálních sítích (altmetriky). U méně významných prací, jako jsou vysokoškolské kvalifikační práce, se nepodařilo prokázat přínos altmetrik pro hodnocení významu eVŠKP. Pro zpracování statistik nad repozitáři eVŠKP v České republice autor doporučuje vybudování centrálního agregačního bodu, který by sbíral a jednotně vyhodnocoval statistiky pomocí jednotlivých metod a standardů popsaných v druhé části kapitoly.

Kapitola 6 obsahuje vypracované doporučení pro výběr discovery služby indexující elektronické informační zdroje, katalog a repozitář eVŠKP univerzity. Cílem doporučení je minimalizace nákladů na zřízení a provoz discovery služby a zároveň eliminování dalších rizik spojených s výběrem a provozem. Uvedené doporučení pro zpracování zadávací dokumentace řeší problematiku nevhodně formulovaných podmínek výběru, které mohou vést až k nutnosti výběrové řízení zrušit. Navržený požadovaný způsob doložení míry pokrytí zdrojů zaručuje jasný a transparentní výpočet, zpětně verifikovatelný. V případě výběrového řízení Univerzity Jana Evangelisty Purkyně v Ústí nad Labem, která si u autora této disertační práce zadala vypracování doporučení, došlo k úspoře přes 3,5 milionu Kč bez DPH oproti původní předpokládané celkové hodnotě zakázky (4).

V kapitole 7. autor navrhuje modelovou implementaci pokročilého vyhledávání metadat a plných textů v repozitáři eVŠKP s metadaty ve formátu EVSKP-MS. Pro potřeby indexování a vyhledávání metadat a plných textů eVŠKP byla autorem disertační práce vybrána platforma Apache Solr. Autor popisuje konkrétní navržený způsob instalace a konfigurace Apache Solr a tvorbu indexu. Import byl otestován na příkladu více jak 30 000 metadatových záznamů a plných textů eVŠKP z Databáze kvalifikačních prací VŠE. V testech vyhledávání provedených autorem se potvrdila velmi rychlá odezva indexovacího serveru. Připravené redesignované uživatelské rozhraní Databáze kvalifikačních prací na bázi

PHP, využívající aplikačního rozhraní Apache Solr pro vyhledávání, umožňuje využití faset a jejich postupné přidávání a odebírání za účelem zpřesnění dotazu.

Rozsáhlá analýza nejvýznamnějších systémů na vyhledávání duplicit u eVŠKP, publikovaná v roce 2013 v rámci 6. ročníku Semináře ke zpřístupňování šedé literatury, kriticky hodnotí jednotlivé systémy v komparaci s prototypovou aplikací autora GooglePlagiarism, využívající vyhledávání pomocí vyhledávače Google. Výsledky analýzy již byly v roce 2014 využity např. při výběru systému kontroly anglických článků pro recenzované časopisy VŠE v Praze nebo při úpravách antiplagiátorských systémů Masarykovy univerzity. Autor disertační práce využívá prezentovanou prototypovou aplikaci GooglePlagiarism mj. při zpracování analýz plagiátorství pro Českou televizi.

Případová studie popsaná v kapitole 9 popisuje vývoj webového portálu Validátor VŠE, který může posloužit jako modelová aplikace zpřístupňování výsledků antiplagiátorské kontroly z externích systémů cílové skupině – vedoucím a oponentům prací. Aplikace je na VŠE v Praze využívána akademickou obcí pro kontrolu eVŠKP a dalších odborných textů v systémech Theses.cz a Odevzdej.cz.

Dílčí závěry a doporučení jsou detailněji formulovány v závěru každé kapitoly.

Kromě výše popsaných přínosů je vhodné zmínit rozsáhlé výsledky více jak desetileté činnosti autora v oblasti správy, vyhledávání a zpřístupňování eVŠKP v ČR. Jako člen a později předseda Komise eVŠKP se významně podílel na přípravě metadatového standardu EVSKP-MS pro popis eVŠKP a jejich komunikaci, na přípravě souvisejících specifikací PersCZ a CorpCZ pro popis fyzických osob a korporací a na přípravě metadatové sady Národního úložiště šedé literatury. Jako řešitel, případně spoluřešitel VŠE v Praze se podílel např. na projektech národního registru VŠKP Theses.cz a Národního úložiště šedé literatury. Je mj. autorem dvou kapitol v recenzovaných knihách, autorsky se spolupodílel na vydání *Metodiky zpracování, dlouhodobého uchování a zpřístupnění šedé literatury v ČR na příkladu Národního úložiště šedé literatury* certifikované Ministerstvem kultury ČR, podílel se na přípravě a pořádání řady odborných seminářů na téma zpřístupňování eVŠKP a šedé literatury. V současné době Jan Mach působí na praxi zpřístupňování eVŠKP např. z pozice člena Výkonného výboru AKVŠ ČR, Rady pro vývoj Národního úložiště šedé literatury nebo Rady DART-Europe.

Seznam literatury a zdrojů

Použitá literatura v textu

Seznam v textu použité literatury je řazen dle prvního výskytu citace v textu disertační práce. Jednotlivé záznamy jsou uvedeny podle normy *ČSN ISO 690 (01 0197)* platné od 1. dubna 2011. Seznam je číslován podle pořadí výskytu.

Při odkazování na očíslovaný seznam literatury je použito metody číselných odkazů v kulatých závorkách.

1. BRATKOVÁ, Eva a Jan MACH. *EVSKP-MS: Metadatový soubor pro elektronické vysokoškolské kvalifikační práce v ČR* [online]. Verze 1.1. Praha: Odborná komise pro otázky elektronického zpřístupňování VŠKP AKVŠ ČR. 15. 7. 2008 [cit. 11. 8. 2014]. Dostupné z: http://www.evskp.cz/standardy/evskp/1.1/

2. BRATKOVÁ, Eva a Jan MACH. *PersCZ: metadatový soubor pro popis fyzických osob* [online]. Praha: Odborná komise pro otázky elektronického zpřístupňování VŠKP AKVŠ ČR. 5. 5. 2008 [cit. 15. 1. 2015]. Dostupné z: http://www.evskp.cz/standardy/perscz/1.0/

3. BRATKOVÁ, Eva a Jan MACH. *CorpCZ: metadatový soubor pro popis korporací* [online]. Praha: Odborná komise pro otázky elektronického zpřístupňování VŠKP AKVŠ ČR. 5. 5. 2008 [cit. 15. 1. 2015]. Dostupné z: http://www.evskp.cz/standardy/corpcz/1.0/

4. UNIVERZITA J. E. PURKYNĚ V ÚSTÍ N. LABEM. *Vyhledávací systém elektronických informačních zdrojů – 2013/0100*. In: ČESKO. Ministerstvo pro místní rozvoj. *Věstník veřejných zakázek* [online]. 11. 12. 2013 [cit. 11. 1. 2015]. Evidenční číslo zakázky: 369400. Dostupné z: http://www.vestnikverejnychzakazek.cz/en/Form/Display/459984

5. ČESKO. Zákon č. 111/1998 sb., o vysokých školách a o změně a doplnění dalších zákonů (zákon o vysokých školách). In: *Sbírka zákonů*. 1998, částka 39, str. 5388 - 5419. ISSN 1211-1244

6. GRULICH, Petr. Vysokoškolské kvalifikační práce jako specifický typ archiválie a jejich digitalizace na Univerzitě Hradec Králové. In: *Konference archivářů České republiky: 11. celostátní konference archivářů České republiky: sborník příspěvků: Chrudim, 4. - 6. května 2005*. Praha: Národní archiv ve spolupráci s Českou archivní společností, 2006. Zpravodaj pobočky České informační společnosti, sv. 49/2005. ISBN 80-8671-2419.

7. ČESKO. Zákon č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon). In: *Sbírka zákonů*. 2000, částka 36, str. 1658 - 1685. ISSN 1211-1244

8. MACH, Jan. Návrh na novelizaci AZ. *Ministerstvo kultury ČR* [online]. 2011 [cit. 3. 1. 2014]. Dostupné z: http://www.mkcr.cz/assets/21--obcan.pdf

9. VOLEMANOVÁ JURMANOVÁ, Věra. Novela vysokoškolského zákona nabyla účinnosti… A co bude dál? *Ikaros* [online]. 2006, roč. 10, č. 2. ISSN 1212-5075. Dostupné z: http://ikaros.cz/novela-vysokoskolskeho-zakona-nabyla-ucinnosti%E2%80%A6-a-co-bude-dal

10. MASARYKOVA UNIVERZITA. Právnická fakulta. *Digitální zpracování tzv. šedé literatury pro Národní úložiště šedé literatury: opinio (stanovisko k právní otázce)* [online]. 30. 7. 2009 [cit. 8. 10. 2014]. Dostupné z: http://invenio.nusl.cz/record/111528/files/idr-284_1.pdf

11. HAVLOVÁ, Jaroslava a Jiří MAREK. Bezplatný otevřený přístup. In: KTD: Česká terminologická databáze knihovnictví a informační vědy (TDKIV) [online]. Praha: Národní knihovna ČR, 2003- [cit. 2015-03-05]. Dostupné z: http://aleph.nkp.cz/F/?func=direct&doc_number=000015853&local_base=KTD

12. CHAN, Leslie aj. Read the Budapest Open Access Initiative. *Budapest Open Access Initiative* [online]. 14. 2. 2002 [cit. 26. 8. 2012]. Dostupné z: http://www.soros.org/openaccess/read

13. Bethesda Statement on Open Access Publishing. *Bethesda* [online]. 20. 6. 2003 [cit. 26. 8. 2012]. Dostupné z: http://www.earlham.edu/~peters/fos/bethesda.htm

14. Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities. *Open Access at the Max Planck Society* [online]. 22. 10. 2003 [cit. 31. 8. 2012]. Dostupné z: http://oa.mpg.de/files/2010/04/berlin_declaration.pdf

15. Budapešťská iniciativa pro Open Access. *Open Access* [online]. 14. 2. 2002 [cit. 5. 3. 2015]. Dostupné z: http://openaccess.cz/cs/iniciativa/

16. *ČSN 5127:2003 Informace a dokumentace - Slovník.* Praha: Český normalizační institut, 2003.

17. Disciplinárním řád pro studenty fakult Vysoké školy ekonomické v Praze. *Vysoká škola ekonomická v Praze* [online]. 26. 4. 1999 [cit. 17. 5. 2013]. Dostupné z: http://www.vse.cz/predpisy/123

18. IPARADIGMS. The Plagiarism Spectrum: Tagging 10 Types of Unoriginal Work. *Turnitin* [online]. 2012 [cit. 7. 7. 2013]. Dostupné z: http://pages.turnitin.com/plagiarism_spectrum.html

19. MACH, Jan. Zpřístupňování vysokoškolských kvalifikačních prací v roce 2014. *ProInflow* [online]. 2014, sv. 6, č. 2. [cit. 2. 2. 2014] Dostupné z: http://www.phil.muni.cz/journals/index.php/proinflow/article/view/1027

20. VYČÍTALOVÁ, Lucie. Aktuální stav zpřístupňování vysokoškolských kvalifikačních prací v ČR. Výsledky průzkumu z října 2009. *Odborná komise pro otázky elektronického zpřístupňování vysokoškolských kvalifikačních prací* [online]. 26. 4. 2009 [cit. 08. 11. 2014]. Dostupné z: http://www.evskp.cz/Dokumentyver/pruzkum2009-100430133025.pdf

21. BALABÁNOVÁ, Tereza. *Problematika zpřístupňování e-VŠKP v rámci Masarykovy univerzity*. Brno, 2006. Bakalářská práce. Masarykova univerzita. Vedoucí práce Věra Jurmanová Volemanová. Dostupné také z: http://is.muni.cz/th/109381/ff_b/

22. ABSOLONOVÁ, Šárka. *Problémy zpřístupňování elektronických vysokoškolských kvalifikačních prací.* Brno, 2008. Bakalářská práce. Masarykova univerzita. Vedoucí práce Petra Šedinová. Dostupné také z: http://theses.cz/id/96svdf/

23. ZLATOHLÁVKOVÁ, Růžena. *Digitální repozitáře na vysokých školách v České republice*. Praha, 2014. Diplomová práce. Univerzita Karlova v Praze. Vedoucí práce Eva Bratková. Dostupné také z: https://is.cuni.cz/webapps/zzp/detail/97157/12746410/

24. BARANAYOVÁ, Irena. *Digitální knihovny kvalifikačních prací v ČR a v zahraničí.* Praha, 2010. Diplomová práce. Univerzita Karlova v Praze. Vedoucí práce Martin Souček. Dostupné také z: https://is.cuni.cz/webapps/zzp/detail/86397/

25. BUGAJEVOVÁ, Jitka. *Dostupnost českých vysokoškolských kvalifikačních prací.* Brno, 2013. Diplomová práce. Masarykova univerzita. Vedoucí práce Petr Škyřík. Dostupné také z: http://is.muni.cz/th/261879/ff_m/

26. HENDL, Jan. *Jak se vyrábí sociologická znalost: příručka pro uživatele.* Praha: Karolinum, 1999. ISBN 978-80-246-1966-8.

27. FAGAN, Jodi Condit. Usability Studies of Faceted Browsing: A Literature Review. *Information Technology and Libraries.* 2010, sv. 29, č. 2, s. 58 - 66. Dostupné také z: http://ejournals.bc.edu/ojs/index.php/ital/article/viewFile/3144/2758

28. MACH, Jan a Iva HOROVÁ. *National Repositories Of ETDs And Grey Literature in Czech Republic.* Poster prezentovaný na ETD2009 conference. University of Pittsburgh. Pittsburgh, 2009. Dostupný také z: http://conferences.library.pitt.edu/ocs/viewpaper.php?id=680&cf=7

29. SULEMAN, Hussein a Edward A. FOX. *Towards Universal Accessibility of ETDs: Building the NDLTD Union Archive.* In: *Proceedings of the ETD 2002*. BYU, Utah, 2002.

30. HICKEY, Thom, Ana PAVANI a Hussein SULEMAN. ETD-MS v1.1: an Interoperability Metadata Standard for Electronic Theses and Dissertations. *NDLTD* [online]. verze 1.1, 19. 8. 2010 [cit. 15. 1. 2015]. Dostupné z: http://www.ndltd.org/standards/metadata

31. LAGOZE, Carl a kol., ed. The Open Archives Initiative Protocol for Metadata Harvesting [online]. Protocol Version 2.0, Document Version 2015-01-08, 14. 6. 2002 [cit. 15. 2. 2015]. Dostupné z: http://www.openarchives.org/OAI/openarchivesprotocol.html

32. CHACHRA, Vinod. *NDLTD Union Catalog / VTLS Visualizer*. In: *ETD 2009 12th International Symposium on Electronic Theses and Dissertations* [online]. 24. 9. 2009 [cit. 10. 11. 2013]. Dostupné z: http://conferences.library.pitt.edu/ocs/viewabstract.php?id=773&cf=7

33. MACH, Jan. *Inspirujeme se v zahraničí*. In: *Systémy pro zpřístupňování VŠKP: zkušenosti, možnosti, nabídky, potřeby* [online prezentace]. Praha: Odborná komise pro otázky elektronického zpřístupňování VŠKP AKVŠ ČR. 2. 11. 2009 [cit. 21. 11. 2013]. ISSN 1803-7003 Dostupné z: http://www.evskp.cz/Seminar4/seminar4-Mach_CZ.ppt

34. History and Milestones. *ProQuest* [online]. 2013 [cit. 23. 11. 2013]. Dostupné z: http://www.proquest.com/en-US/aboutus/history.shtml

35. ProQuest® Dissertations & Theses. *ProQuest* [online]. 2013 [cit. 23. 11. 2013]. Dostupné z: http://www.proquest.com/assets/literature/products/databases/pqdt.pdf

36. ROSS, Amanda a Vladimír KAREN. *Uživatelský průzkum ProQuest Dissertations & Theses: Jak pracují uživatelé s informacemi?* In: *Systémy pro zpřístupňování VŠKP: zkušenosti, možnosti, nabídky, potřeby* [online prezentace]. Praha: Odborná komise pro otázky elektronického zpřístupňování VŠKP AKVŠ ČR. 2. 11. 2009. [cit. 20. 11. 2013] ISSN 1803-7003 Dostupné z: http://www.evskp.cz/Seminar4/seminar4-Karen_CZ.ppt

37. THE BRITISH LIBRARY BOARD. About EThOS. *Electronic Theses Online Service* [online]. The British Library, 2013 [cit. 20. 11. 2013]. Dostupné z: http://ethos.bl.uk/About.do

38. EThOS Toolkit. *EThOS Toolkit* [online]. The British Library, 2013 [cit. 19. 11. 2013]. Dostupné z: http://cclib-2.dmz.cranfield.ac.uk/ethostoolkit

39. The EThOS UKETD_DC application profile. *EThOS Toolkit* [online]. The British Library, 2013 [cit. 20. 11. 2013]. Dostupné z: http://cclib-2.dmz.cranfield.ac.uk/ethostoolkit/tiki-index.php?page=The+EThOS+UKETD_DC+application+profile

40. MOYLE, Martin. Improving Access to European E-theses: the DART-Europe Programme. *Liber Quarterly*. 2008, Sv. 18, č. 3/4. e-ISSN 2213-056X. Dostupné z: http://liber.library.uu.nl/index.php/lq/article/view/7940/8210

41. DART-Europe Partnership Agreement . *DART-Europe* [online]. Rev. 4. 2. 2008 [cit. 9. 11. 2013]. Dostupné z: http://www.dart-europe.eu/About/documents/DART-Europe_Partnership_Agreement.pdf

42. MOYLE, Martin. *DART-Europe E-theses Portal.* In: *Systémy pro zpřístupňování VŠKP: zkušenosti, možnosti, nabídky, potřeby* [online]. Praha: Odborná komise pro otázky elektronického zpřístupňování VŠKP AKVŠ ČR. 2. 11. 2009. ISSN 1803-7003 Dostupné z: http://www.evskp.cz/Seminar4/seminar4-Moyle-text.pdf

43. DART-EUROPE. *DART-Europe E-theses Portal* [online]. 2015 [cit. 22. 2. 2015]. Dostupné z: http://www.dart-europe.eu/

44. LOCHMAN, Martin. Portál elektronických disertací DART-Europe. *Ikaros* [online]. 2014, roč. 18, č. 5. [cit. 22. 1. 2015] urn:nbn:cz:ik-14227. ISSN 1212-5075. Dostupné z: http://ikaros.cz/node/14227

45. VYSOKÁ ŠKOLA EKONOMICKÁ V PRAZE. Databáze kvalifikačních prací VŠE [aplikace]. 2006 -. Dostupné z: https://www.vse.cz/vskp/

46. Soubor doporučení. *Odborná komise pro otázky elektronického zpřístupňování vysokoškolských kvalifikačních prací* [online]. 2006 [cit. 10. 12. 2013]. Dostupné z: http://www.evskp.cz/dokumenty.php?tsekce=2&sek=&ukol=1

47. ČESKO. Ministerstvo kultury. Stanovisko Samostatného oddělení autorského práva Ministerstva kultury k právnímu názoru odboru legislativního a právního Ministerstva školství, mládeže a tělovýchovy k aplikaci § 47b zákona o vysokých školách č. 111/1998 Sb. *Informace pro knihovny* [online]. 26. 4. 2007 [cit. 10. 12. 2013]. Dostupné z: http://knihovnam.nkp.cz/sekce.php3?page=03_Leg/01_LegPod/Stanovisko111_98.htm

48. UNIVERZITA KARLOVA V PRAZE. Evidence a zveřejňování závěrečných prací. *Univerzita Karlova* [online]. 12. 11. 2014 [cit. 16. 2. 2015]. Dostupné z: http://www.cuni.cz/UK-4474.html

49. BRANDEJS, Michal aj. Prohlášení k centralizovaným rozvojovým projektům řešícím problematiku vysokoškolských kvalifikačních prací. *Odborná komise pro otázky elektronického zpřístupňování vysokoškolských kvalifikačních prací* [online]. 12. 9. 2007 [cit. 29. 11. 2013]. Dostupné z: http://www.evskp.cz/Dokumentyver/prohlaseni.pdf

50. BRANDEJSOVÁ, Jitka a Jan MACH. Projekt NR VŠKP a systém na odhalování plagiátů. In: *Systémy pro zpřístupňování VŠKP: zkušenosti, možnosti, nabídky, potřeby: 2. ročník semináře konaného 16. 10. 2007 na VUT v Brně* [online prezentace]. Praha: Odborná komise pro otázky elektronického zpřístupňování VŠKP AKVŠ ČR. 25. 10. 2007 [cit. 9. 12. 2013]. ISSN 1803-7003. Dostupné z: http://www.evskp.cz/Seminar2/seminar2-brandejsova.pdf

51. NÁRODNÍ TECHNICKÁ KNIHOVNA. Konverzní tabulka mezi NUŠL formátem a MARC 21. *Národní úložiště šedé literatury* [online]. 7. 12. 2010 [cit. 8. 12. 2014]. Dostupné z: http://nusl.techlib.cz/images/Konverzn%C3%AD_tabulka_NU%C5%A0L_MARC_21.pdf

52. HOROVÁ, Iva a Jarmila KRKOŠKOVÁ. *Aktuální stav zpřístupňování vysokoškolských kvalifikačních prací v ČR: výsledky průzkumu* [online]. Odborná komise pro otázky elektronického zpřístupňování VŠKP AKVŠ ČR. 14. 3. 2007 [cit. 8. 11. 2014]. Dostupné z: http://www.evskp.cz/Dokumentyver/vyzkum-070622150034.pdf

53. HOROVÁ, Iva. *Aktuální stav zpřístupňování vysokoškolských kvalifikačních prací v ČR: výsledky průzkumu z prosince 2007* [online]. Odborná komise pro otázky elektronického zpřístupňování VŠKP AKVŠ ČR. 6. 3. 2008 [cit. 8. 11. 2014]. Dostupné z: http://www.evskp.cz/Dokumentyver/pruzkum2007-080306104121.pdf

54. ČESKO. Ministerstvo školství, mládeže a tělovýchovy. *Veřejné vysoké školy* [online]. 2014 [cit. 2. 12. 2014]. Dostupné z: http://www.msmt.cz/vzdelavani/vysoke-skolstvi/verejne-vysoke-skoly-4

55. MASARYKOVA UNIVERZITA. Seznam správců. *Theses.cz: Vysokoškolské kvalifikační práce* [online]. 2014 [cit. 10. 11. 2014]. Dostupné z: http://theses.cz/spravci/

56. MASARYKOVA UNIVERZITA. Co umí náš systém? *Informační systém Masarykovy univerzity* [online]. 2014 [cit. 15. 11. 2014] Dostupné z: http://is.muni.cz/nas_system/moznosti.pl

57. IS4U. UIS Univerzitní informační systém: z pohledu jeho uživatelů. *UIS Univerzitní informační systém* [online]. 2014 [cit. 29. 11. 2014]. Dostupné z: http://www.uis-info.com/__files/uis-web/documents/brozura_uis_uzivatele_cz_small.pdf

58. ZÁPODOČESKÁ UNIVERZITA. IS/STAG: Informační systém studijní agendy. *IS/STAG* [online]. 4. 2. 2013 [cit. 28. 11. 2014]. Dostupné z: http://is-stag.zcu.cz/

59. BRATKOVÁ, Eva. *Digitální knihovny s volným přístupem v oblasti vědy a výzkumu a identifikace a metadatový popis jejich objektů*. Praha, 2009. Disertační práce. Univerzita Karlova v Praze. Vedoucí práce Rudolf Vlasák. Dostupné také z: https://is.cuni.cz/webapps/zzp/detail/75005/

60. GODFREY, Neil. Electronic Theses and Dissertation Metadata Schema (ETD-MS) for Australia? In: *Metalogger* [online]. 2. 7. 2007. [cit: 7. 6. 2014]. Dostupné z: https://metalogger.wordpress.com/2007/07/02/electronic-theses-and-dissertation-metadata-schema-ETD-MS-for-australia/

61. DCMI USAGE BOARD. Dublin Core Metadata Element Set: version 1.1. *DCMI* [online]. 14. 6. 2012 [cit. 5. 12. 2014]. Dostupné z: http://dublincore.org/documents/dces/

62. DRIVER. Pokyny k DRIVER 2.0. *NTK v Praze* [online]. 2008 [cit. 9. 11. 2013]. Dostupné z: http://www.techlib.cz/files/download/id/2476/driver-guidelines.pdf

63. OpenAIRE Guidelines: For Literature repositories. *OpenAIRE Guidelines wiki* [online]. 3. 4. 2013 [cit. 18. 12. 2014]. Dostupné z: https://guidelines.openaire.eu/wiki/OpenAIRE_Guidelines:_For_Literature_repositories

64. DCMI USAGE BOARD. DCMI Metadata Terms. *DCMI* [online]. 14. 6. 2012 [cit. 11. 12. 2014]. Dostupné z: http://dublincore.org/documents/2012/06/14/dcmi-terms/

65. MASARYKOVA UNIVERZITA. Formát importu dat theses.cz. *Theses.cz: Vysokoškolské kvalifikační práce* [online]. 2008 [cit. 15. 1. 2015]. Dostupné z: https://theses.cz/auth/th_dok/format_theses_10.pl

66. MACH, Jan. Přenos VŠKP pomocí protokolu OAI-PMH. In: *Systémy pro zpřístupňování VŠKP: zkušenosti, možnosti, nabídky, potřeby* [online]. Praha: Odborná komise pro otázky elektronického zpřístupňování VŠKP AKVŠ ČR. 7. 10. 2008 [cit. 9. 12. 2013]. ISSN 1803-7003. Dostupné z: http://www.evskp.cz/Seminar3/seminar3-Mach2a.pdf

67. MACH, Jan. *Metriky pro open access repozitáře.* Praha*,* 2012. Atestační práce z předmětu Informační věda. Univerzita Karlova.

68. LOZANO, George A. a Vincent LARIVIÈRE, Yves GINGRAS. The weakening relationship between the Impact Factor and papers' citations in the digital age. *Journal of the American Society for Information Science and Technology*. 8. 10. 2012. DOI: 10.1002/asi.22731. Dostupné také z: http://arxiv.org/abs/1205.4328

69. IVANCHEVA, Ludmila. Scientometrics Today: a Methodological Overview. In: *Collnet journal of scientometrics and information management*. 12. 2008, sv. 2, č. 2. ISSN: 0973-7766

70. GARFIELD, Eugene. The History and Meaning of the Journal Impact Factor. *JAMA. 4. 1.* 2006, sv. 293, č. 1, s. 90-93. Dostupné také z: http://garfield.library.upenn.edu/papers/jamajif2006.pdf

71. GARFIELD, Eugene. Journal impact factor: a brief review. *Canadian Medical Association Journal.* 1999, sv. 161, č. 8, s. 979-980. Dostupné také z: http://www.garfield.library.upenn.edu/papers/journalimpactCMAJ1999.pdf

72. VAVŘÍKOVÁ, Lucie. *Úvod do scientometrie* [online]. 2008 [cit. 31. 8. 2012]. Dostupné z: http://tarantula.ruk.cuni.cz/KPSV-9-version1-scientometrie.pdf

73. AMIN, Mayur a Michael MABE. ImpactFactors: Use and Abuse. *Perspectives in Publishing.* 2000, č. 1, s. 1-6. Reissued with minor revisions 11. 2007. Dostupné z: http://cdn.elsevier.com/assets/pdf_file/0014/111425/Perspectives1.pdf

74. GARFIELD, Eugene. Whither Journals and Impact Factors? In: BRAUN, T., ed. *Impact Factor of Scientific and Scholarly Journals. Its Use and Misuse in Research Evaluation.* Budepest: Akademiai Kiado, 2007, str. v-vi. Scientometrics Guidebooks Series - Volume 2. Dostupné také z: http://garfield.library.upenn.edu/papers/whitherjournalsimpactfactor2007.pdf

75. HERB, Ulrich. Alternative Impact Measures for Open Access Documents? An examination how to generate interoperable usage information from distributed open access services. In: *Open access to knowledge - promoting sustainable progress. World Library and information congress: 76th IFLA general confrence and assembly: 10.-15. 8. 2010, Gothenburg,* 2010. Dostupné také z: http://conference.ifla.org/past-wlic/2010/72-herb-en.pdf

76. ROUSSEAU, Ronald. New developments related to the Hirsch index [online]. [cit. 31. 8. 2012]. Preprint. Dostupné z: http://sci2s.ugr.es/impact/Hirsch_new_developments.pdf

77. EGGHE, Leo. Theory and practise of the g-index. *Scientometrics.* 2006, Sv. 69, č. 1, s. 131-152

78. BRODY, Tim, CARR, Les, GINGRAS, Yves, HAJJEM, Chawki, HARNARD, Stevan, SWAM, Alma, DIRKS, Lee a HEY, Tony (ed.). Incentivizing the Open Access Research Web: publication-Archiving, Data-Archiving and Scientometrics. *CTWatch Quarterly*, 2007, sv. 3, č. 3. Dostupné také z: http://eprints.soton.ac.uk/264418/

79. Erhebungshandbuch für den BIX für Öffentliche Bibliotheken 2012 (Berichtsjahr 2011). *BIX* [online]. 29. 11. 2011 [cit. 3. 8. 2012]. Dostupné z: http://www.bix-bibliotheksindex.de/fileadmin/user_upload/Projektinfos/Erhebungsunterlage_BIX-OEB_2012.pdf

80. SCHOLZE, Frank aj. Briefing Paper: Combined usage statistics as a basis for Research intelligence. *Knowledge Exchange* [online]. 30. 3. 2009 [cit. 5. 8. 2012] Dostupné z: http://www.knowledge-exchange.info/Default.aspx?ID=393

81. Counter Online Metrics: The COUNTER Code of Practice for e-Resources: Release 4. *COUNTER* [online]. 4. 2012 [cit. 11. 8. 2012]. Dostupné z: http://www.projectcounter.org/r4/COPR4.pdf

82. NATIONAL INFORMATION STANDARDS ORGANIZATION (U.S.) a AMERICAN NATIONAL STANDARDS INSTITUTE. *The Standardized Usage Statistics Harvesting Initiative (SUSHI) protocol: an American national standard*. Baltimore, Md.: National Information Standards Organization, 29. 10. 2007. National information standards series. ISBN 9781880124703 188012470X.

83. COUNTER. COUNTER Code of Practice for Articles. In: *Counter Online Metrics* [online]. 4. 2012 [cit. 5. 12. 2014]. Dostupné z: http://www.projectcounter.org/documents/counterart_cop_MAR2014.pdf

84. PRIEM, Jason aj. Altmetrics: a manifesto. In: *Altmetrics* [online]. Verze 1.01, 28. 9. 2011 [cit. 3. 8. 2012]. Dostupné z: http://altmetrics.org/manifesto/

85. The OpenAIRE Consortium. Usage Statistics from repositories. *OpenAire: open access infrastructure for reserach in Europe* [online]. 2012 [cit. 5. 2. 2015]. Dostupné z: https://www.openaire.eu/content

86. PRIEM, Jason, Heather A. Piwowar a Bradley M. Hemminger. Altmetrics in the Wild: Using Social Media to Explore Scholarly Impact. In: *ACM Web Science Conference 2012 Workshop*. Evanston, IL, 21. 6. 2012 [cit. 31. 8. 2012]. Dostupné také z: http://arxiv.org/abs/1203.4745v1

87. *PLOS ONE* [online]. San Francisco: PLOS, 2006- [cit. 13. 8. 2012]. eISSN 1932-6203 Dostupné z: http://www.plosone.org/

88. SHEPHERD, Peter a Paul NEEDHAM. *Publisher and Institutional Repository usage Statistics: The PIRUS2 Project: final report* [online]. Cranfield: Cranfield University, 6. 10. 2011 [cit. 13. 8. 2012]. Dostupné z: http://www.cranfieldlibrary.cranfield.ac.uk/pirus2/tiki-download_wiki_attachment.php?attId=170&download=y

89. VERHAAR, Peter. KE Usage Statistics Guidelines. *SUFR wiki* [online]. Verze 1.0, 18. 5. 2010 [cit. 15. 8. 2012]. Dostupné z: http://wiki.surf.nl/display/standards/KE+Usage+Statistics+Guidelines

90. IRUS-UK. *IRUS-UK* [online]. 2014 [cit. 1. 8. 2014]. Dostupné z: http://www.irus.mimas.ac.uk/

91. NEEDHAM, Paul. IRUS UK: Making ETDs count in UK repositories. In: *ETD 2014*. University of Lecester, 2014. Dostupné také z: https://www.youtube.com/watch?v=flocun3wAZU

92. The PIRUS Code of Practice for recording and reporting usage at the individual article level. In: *PIRUS* [online]. Verze 1. říjen 2013 [cit. 5. 9. 2014]. A COUNTER Standard. Dostupné z: http://www.projectcounter.org/documents/Pirus_cop_OCT2013.pdf

93. The Tracker protocol V3.1. *IRUS-UK* [online]. 22. 4. 2014 [cit. 5. 9. 2014]. Dostupné z: http://www.irus.mimas.ac.uk/help/toolbox/TrackerProtocol-V3-2014-04-22.pdf

94. Homepage. PLUM ANALYTICS. *PlumX* [online]. 2014 [cit. 8. 8. 2014]. Dostupné z: https://plu.mx/

95. UNIVERZITA PARDUBICE. Dodávka discovery systému. In: ČESKO. Ministerstvo pro místní rozvoj. *Věstník veřejných zakázek* [online]. 21. 10. 2011 [cit. 8. 11. 2014]. Evidenční číslo zakázky: 60066836. Dostupné z: http://vestnikverejnychzakazek.cz/cs/Form/Display/26966

96. UNIVERZITA PARDUBICE. Poskytnutí softwarové licence na discovery systém. In: ČESKO. Ministerstvo pro místní rozvoj. *Věstník veřejných zakázek* [online]. 9. 8. 2012 [cit. 8. 11. 2014]. Evidenční číslo zakázky: 210909. Dostupné z: http://vestnikverejnychzakazek.cz/cs/Form/Display/351163

97. NÁRODNÍ TECHNICKÁ KNIHOVNA. Univerzální vyhledávač elektronických informačních zdrojů. In: ČESKO. Ministerstvo pro místní rozvoj. *Věstník veřejných zakázek* [online]. 7. 12. 2012 [cit. 7. 11. 2014]. Evidenční číslo zakázky: 240227. Dostupné z: http://vestnikverejnychzakazek.cz/cs/Form/Display/375297

98. SIKORA, Radek. *Vyhledávání v českých dokumentech pomocí Apache Solr.* Brno, 2012. Diplomová práce. Masarykova univerzita. Vedoucí práce Radek Ošlejšek. Dostupné také z: http://is.muni.cz/th/256499/fi_m/

99. PETYLKA, Petr. *Jak kvalita lemmatizace ovlivňuje výsledky vyhledávání dokumentů v českém jazyce.* Praha, 2012. Vysoká škola ekonomická v Praze. Vedoucí práce Petr Strossa. Dostupné také z: https://www.vse.cz/vskp/34929_jak_kvalita_lemmatizace_ovlivnuje_vysledky_vyhledavani_dokumentu_v%C2%A0ceskem_jazyce

100. THE APACHE SOFTWARE FOUNDATION. Solr Quick Start. *Solr* [online]. 2014 [cit. 13. 3. 2014]. Dostupné také z: http://lucene.apache.org/solr/quickstart.html

101. SCIRUS ETD Search. *Scirus* [online]. 2013 [cit. 24. 4. 2013]. Dostupné z: http://www.ndltd.org/serviceproviders/scirus-etd-search

102. ČIČÁK, Matěj. Google vs. Seznam: skóre je 5:3, odhalil průzkum. *Živě.cz* [online]. 28. 2. 2013 [cit. 12. 4. 2013]. Dostupné z: http://www.zive.cz/clanky/google-vs-seznam-skore-je-53-odhalil-pruzkum/sc-3-a-167776/

103. MICROSOFT. *Microsoft Translator* [online]. 2013 [cit. 17. 4. 2013]. Dostupné z: http://www.microsoft.com/en-us/translator/

104. IPARADIGMS. Our Company. *Turnitin* [online]. 2013 [cit. 15. 5. 2013]. Dostupné z: http://turnitin.com/en_us/about-us/our-company

105. This is how it works. *Ephorus* [online]. 2013 [cit. 28. 6. 2013]. Dostupné z: https://www.ephorus.com/this-is-how-it-works/

106. Bohuňovská opsala diplomovou práci, říká expert na plagiátorství. *Parlamentní listy* [online]. 20. 5. 2011 [cit. 26. 6. 2013]. Dostupné z: http://www.parlamentnilisty.cz/arena/monitor/Bohunovska-opsala-diplomovou-praci-rika-expert-na-plagiatorstvi-197771

107. MACH, Jan. Brněnský plagiát… In: *Události, komentáře.* TV, ČT 24, 20. 5. 2011 22:21. Dostupný také z: http://www.ceskatelevize.cz/ivysilani/1096898594-udalosti-komentare/211411000370520/obsah/158009-brnensky-plagiat

108. MACH, Jan. Nepůvodní diplomová práce. In: *Reportéři ČT*. TV, ČT 1. 9. 5. 2011 21:25. Dostupné také z: http://www.ceskatelevize.cz/ivysilani/1142743803-reporteri-ct/211452801240018

109. GOOGLE. Custom Search: Overview. *Google Developers* [online]. 3. 8. 2012 [cit. 22. 2. 2015]. Dostupné z: https://developers.google.com/custom-search/v1/overview?hl=cs

110. ČESKO. Ministerstvo školství, mládeže a tělovýchovy. *Vyhlášení rozvojových programů pro rok 2013* [online]. 10. 10. 2012 [cit. 14. 10. 2012]. Dostupné z: http://www.msmt.cz/file/24657

111. MASARYKOVA UNIVERZITA. Detail projektu Meziuniverzitní síť technických a metodických opatření na ochranu proti plagiátorství [online]. 2012 [cit. 14. 10. 2012]. Dostupné z: http://www.muni.cz/research/projects/15563

112. VYSOKÁ ŠKOLA EKONOMICKÁ V PRAZE. *Validátor VŠE* [aplikace]. 2011 -. Dostupné z: http://validator.vse.cz

113. JASSO. How to remove namespaces from XML using XSLT. *StackOverflow* [online]. 3. 5. 2011 [cit. 18. 3. 2014]. Dostupné z: http://stackoverflow.com/questions/5268182/how-to-remove-namespaces-from-xml-using-xslt

114. PROQUEST. ProQuest Dissertations and Master's Theses Publishing [Jak publikovat dizertace zdarma v PQDT]. *Albertina icome Praha* [online]. 2011 [cit. 27. 11. 2013]. Dostupné z: http://www.aip.cz/podpora/nastroje/345-jak-publikovat-dizertace-zdarma-v-pqdt/

115. THELWALL, Mike, Liwen VAUGHAN a Lennart BJÖRNEBORN. Webometrics. *Annual Review of Information Science and Technology.* 2005, sv. 39, č. 1, stránky 81-135. DOI: 10.1002/aris.1440390110

116. BLAŽEJ, Josef. Hirschův index. *Wikimedia Commons* [online]. 2. 12. 2008 [cit. 25. 7. 2012]. Dostupné z: http://commons.wikimedia.org/wiki/File:Hirschuv_index.png

117. MACH, Jan. Zpřístupnění vysokoškolských kvalifikačních prací. In: PEJŠOVÁ, Petra. *Repozitáře šedé literatury*. Zlín: VeRBuM, 2010, s. 55-65. ISBN 978-80-904273-5-8. Dostupné také z: http://nusl.techlib.cz/images/Book.pdf

118. NÁRODNÍ TECHNICKÁ KNIHOVNA. Softwarové řešení NUŠL. *Národní úložiště šedé literatury* [online]. 2008 [cit. 9. 11. 2014]. Dostupné z: http://nusl.techlib.cz/index.php/Software

119. MACH, Jan. A Comparison Of Anti-Plagiarism Systems For Theses And Dissertations. In: *Seminář ke zpřístupňování šedé literatury 2013: 6. ročník semináře zaměřeného na problematiku uchovávání a zpřístupňování šedé literatury* [online]. Praha: Národní technická knihovna. 23. 10. 2013 [cit. 5. 12. 2014]. ISSN 1803-6015. Dostupné z: http://nrgl.techlib.cz/images/Mach_fulltext.pdf

120. IMPACTSTORY. Carl Boettiger: Is your phylogeny informative? Measuring the power of comparative methods. *Impactstory* [online]. 2015 [cit. 18. 2. 2015]. Dostupné z: https://impactstory.org/CarlBoettiger/product/t2q1a39jt3kythditpt30uhu/metrics

Publikační činnost autora

Výsledky výzkumu byly odborné veřejnosti autorem disertační práce prezentovány v uplynulých letech formou grantových výstupů, případových studií realizovaných implementací a v publikacích autora (např. kapitoly v knize, prezentace, reportáže ve veřejnoprávní televizi, certifikovaná metodika, metadatové standardy). Níže je uveden výběr z publikační činnosti autora v oblasti správy, vyhledávání a zpřístupňování elektronických vysokoškolských kvalifikačních prací a zpracovaných analýz.

Publikace a prezentace

MACH, Jan. Statistiky využití článků v online repozitářích. In: *INFORUM 2015: 21. ročník konference o profesionálních informačních zdrojích, Praha 26.-27. 5. 2015* [online]. Praha: Albertina icome Praha, 2014. ISSN 1801–2213. Dostupný z WWW: http://www.inforum.cz/cs/sbornik/
V březnu 2015 byl přijat příspěvek Jana Macha „Statistiky využití článků v online repozitářích" na konferenci INFORUM 2015, která se bude konat v květnu 2015. Příspěvek bude následně zveřejněn ve sborníku.

MACH, Jan. Zpřístupňování vysokoškolských kvalifikačních prací v roce 2014. *ProInflow* [online]. Brno: Masarykova univerzita, 2014, sv. 6, č. 2 [cit. 2. 2. 2014]. Dostupné z: http://www.phil.muni.cz/journals/index.php/proinflow/article/view/1027
Článek v recenzovaném časopise ProInflow k realizovanému průzkumu Zpřístupňování vysokoškolských kvalifikačních prací v roce 2014.

MACH, Jan. Repozitáře vysokoškolských kvalifikačních prací [webinář]. Praha: Univerzita Karlova, 15. 10. 2014 [cit. 2. 2. 2014]. Dostupné z: https://connect.cesnet.cz/p8dnk15599o/
Webinář přednesený v rámci projektu Rozvoj portálu Elektronických studijní textů z oboru ICT, informační vědy a knihovnictví (program VISK 2, podprogram - Mimoškolní vzdělávání knihovníků).

MACH, Jan. A Comparison Of Anti-Plagiarism Systems For Theses And Dissertations. In: *Seminář ke zpřístupňování šedé literatury 2013: 6. ročník semináře zaměřeného na problematiku uchovávání a zpřístupňování šedé literatury* [online]. Praha: Národní technická knihovna, 23. 10. 2013 [cit. 5. 12. 2014]. ISSN 1803-6015. Dostupné z: http://nrgl.techlib.cz/images/Mach_fulltext.pdf
Jan Mach je členem programového výboru semináře, autorem české a anglické prezentace a příspěvku ve sborníku. Přednáška v evaluaci semináře skončila na druhém místě v hodnocení nejzajímavějších přednášek.

MACH, Jan a Jiří PAVLÍK. Zabezpečení e-knih. In: BOUDA, Tomáš. *Elektronické knihy v českých knihovnách*. Brno: Masarykova univerzita, 2012, s. 43-52. ISBN 978–80–210–6000–5. Dostupné také z: http://eknihy.knihovna.cz/kniha/elektronicke-knihy-v-ceskych-knihovnach
Jan Mach je spoluautorem kapitoly Zabezpečení e-knih, především části týkající se formátu PDF.

MACH, Jan. Akademické výsledky na veřejnosti - hrozba či příležitost? In: *Jinonické informační pondělky* [prezentace]. Praha: ÚISK, 14. 3. 2011. Dostupné také z: http://uisk.ff.cuni.cz/detail.do?articleId=15111
Prezentace v rámci semináře Jinonické informační pondělky seznámila posluchače se stavem zpřístupňování studentských kvalifikačních prací, šedé literatury a výsledků vědy, výzkumu a inovací na vysokých školách včetně přístupu Open Access.

PEJŠOVÁ, Petra (ed.). *Metodika zpracování, dlouhodobého uchování a zpřístupnění šedé literatury v ČR na příkladu Národního úložiště šedé literatury* [online]. Praha: Národní technická knihovna, 2011 [cit. 1. 2. 2015]. Dostupné z: http://nusl.techlib.cz/index.php/Metodika
Jan Mach se autorsky podílel na přípravě části týkající se eVŠKP, metadat a OAI-PMH, jako interní oponent kapitoly Persistentní identifikátor.

MACH, Jan. Zpřístupnění vysokoškolských kvalifikačních prací. In: PEJŠOVÁ, Petra (ed.). *Repozitáře šedé literatury*. Zlín: VeRBuM, 2010, s. 55-65. ISBN 978-80-904273-5-8. Dostupné také z: http://nusl.techlib.cz/images/Book.pdf
Jan Mach je autorem kapitoly Zpřístupnění vysokoškolských kvalifikačních prací. V rámci projektu MK ČR zajišťoval vydání publikace.

MACH, Jan a Iva HOROVÁ. *National Repositories Of ETDs And Grey Literature in Czech Republic.* Poster prezentovaný na ETD2009 conference. University of Pittsburgh. Pittsburgh, 2009. Dostupný také z: http://conferences.library.pitt.edu/ocs/viewpaper.php?id=680&cf=7
Jan Mach je spoluautorem posteru, který osobně prezentoval v roce 2009 na konfernci ETD2009 v Pittsburghu, USA.

BRATKOVÁ, Eva a Jan MACH. *EVSKP-MS: Metadatový soubor pro elektronické vysokoškolské kvalifikační práce v ČR* [online]. Verze 1.1. Praha: Odborná komise pro otázky elektronického zpřístupňování VŠKP AKVŠ ČR. 15. 7. 2008 [cit. 11. 8. 2014]. Dostupné z: http://www.evskp.cz/standardy/evskp/1.1/
Autor se významnou měrou podílel na přípravě standardu EVSKP-MS pro popis eVŠKP, především na reprezentaci jednotlivých metadat ve standardu XML a na návrhu technických a administrativních metadatových prvků.

BRATKOVÁ, Eva a Jan MACH. *CorpCZ: metadatový soubor pro popis korporací* [online]. Praha: Odborná komise pro otázky elektronického zpřístupňování VŠKP AKVŠ ČR. 5. 5. 2008 [cit. 15. 1. 2015]. Dostupné z: http://www.evskp.cz/standardy/corpcz/1.0/
Jan Mach se spolupodílel na návrhu metadatových prvků pro popis korporací v rámci Odborné komise pro otázky elektronického zpřístupňování VŠKP AKVŠ ČR, které předsedal.

BRATKOVÁ, Eva a Jan MACH. *PersCZ: metadatový soubor pro popis fyzických osob* [online]. Praha: Odborná komise pro otázky elektronického zpřístupňování VŠKP AKVŠ ČR. 5. 5. 2008 [cit. 15. 1. 2015]. Dostupné z: http://www.evskp.cz/standardy/perscz/1.0/
Jan Mach se spolupodílel na návrhu metadatových prvků pro popis fyzických osob v rámci Odborné komise pro otázky elektronického zpřístupňování VŠKP AKVŠ ČR, které předsedal.

BRATKOVÁ, Eva a Jan MACH. Metadatový standard EVSKP-MS, verze 1.1 pro popis VŠKP a standardy související. In: *Systémy pro zpřístupňování VŠKP: zkušenosti, možnosti, nabídky, potřeby* [online]. Praha: Odborná komise pro otázky elektronického zpřístupňování VŠKP AKVŠ ČR. 7. 10. 2008 [cit. 25. 12. 2014]. ISSN 1803-7003 Dostupné z: http://www.evskp.cz/Seminar3/seminar3-BratkovaMach.pdf
Jan Mach je spoluautorem příspěvku na semináři, který pořádal, a autorem související prezentace.

BRATKOVÁ, Eva a Jan MACH. Standardizace metadat pro národní registr EVŠKP. In: *MIKULECKÁ, J.; SEDLÁČEK, J. (ed.). Digitální knihovny: provoz a budování: sborník konference EUNIS-CZ* [online]. Špindlerův Mlýn, 28. až 30. 5 2006. Vyd. 1. s přílohou CD. Hradec Králové: Gaudeamus, 2006, s. 5-12. ISBN 80-7041-200-3. Dostupný také z: http://eprints.rclis.org/7754/.
Jan Mach je spoluautorem příspěvku, především návrhu doporučené možnosti zápisu metadat ve formátu RDF/XML, XML a HTML/XHTML a zápis obsahu některých prvků.

MACH, Jan. Možnosti spolupráce: e-VŠKP v rámci Systému šedé literatury. In: *Celostátní porada vysokoškolských knihoven 9. - 10. 11. 2005* [prezentace]. Praha: Česká zemědělská univerzita v Praze, 2005.

MACH, Jan. Návrh národního souboru metadat pro popis e-VŠKP. Reprezentace metadat. In: *Celostátní porada vysokoškolských knihoven 9. - 10. 11. 2005* [prezentace]. Praha: Česká zemědělská univerzita v Praze.

Zpracované analýzy

MACH, Jan. Problémy bakalářů brněnské MU. In: *Události*. TV, ČT 1, 9. 2. 2012 19:00 [cit. 2. 2. 2015]. Dostupný také z: http://www.ceskatelevize.cz/ivysilani/1097181328-udalosti/212411000100209/obsah/189605-problemy-bakalaru-brnenske-mu/
Reportáž s Janem Machem v pořadu Události ke zpracovaným analýzám závěrečných prací.

MACH, Jan. Brněnský plagiát… In: *Události, komentáře.* TV, ČT 24, 20. 5. 2011 22:21. Dostupný také z: http://www.ceskatelevize.cz/ivysilani/1096898594-udalosti-komentare/211411000370520/obsah/158009-brnensky-plagiat
Rozhovor s Janem Machem v pořadu Události, komentáře ke zpracované analýze diplomové práce Jany Bohuňovské.

MACH, Jan. Opisování náměstkyně brněnského primátora. In: *Události*. TV, ČT 1, 20. 5. 2011 19:00 [cit. 2013-12-30]. Dostupný také z: http://www.ceskatelevize.cz/ivysilani/1097181328-udalosti/211411000100520/
Reportáž s Janem Machem v pořadu Události ke zpracované analýze diplomové práce Jany Bohuňovské.

MACH, Jan. Nepůvodní diplomová práce. In: *Reportéři ČT*. TV, ČT 1. 9. 5. 2011 21:25. Dostupné také z: http://www.ceskatelevize.cz/ivysilani/1142743803-reporteri-ct/211452801240018
Reportáž s Janem Machem v pořadu Reportéři ČT ke zpracované analýze diplomové práce Jitky Wiszczorové.

MACH, Jan. *Specifikace vyhledávacího nástroje nové generace (tzv. discovery system)*. Praha, 2013
Příprava podkladů (Definice pojmů, Předmět veřejné zakázky, Minimální technické parametry, Kritérium Pokrytí zdrojů a jeho vyhodnocování, Kritérium Funkce systému a podpora, subkritéria a vyhodnocování) pro výběrové řízení na vyhledávací nástroj nové generace. Zpracováno na zakázku české vysoké školy.

MACH, Jan. *Posudek závěrečných kvalifikačních prací*. Praha, 2013
Posudek rigorózní a diplomové práce JUDr. Radka Ondruše. Obsahová analýza tištěných prací za využití dokumentů v archivu Masarykovy univerzity v Brně. Zpracováno na zakázku právní kanceláře.

MACH, Jan. Diplomové práce z plzeňské fakulty práv. In: *Reportéři ČT*. TV, ČT 1, 9. 11. 2009 21:35. Dostupný také z: http://www.ceskatelevize.cz/porady/1142743803-reporteri-ct/209452801240042/
Reportáž s Janem Machem v pořadu Reportéři ČT ke zpracované analýze diplomové práce Ivany Řápkové.

Příloha I. Prohlášení k centralizovaným rozvojovým projektům řešícím problematiku vysokoškolských kvalifikačních prací

Na základě výzvy České konference rektorů ze dne 6. 9. 2007 se uskutečnilo společné jednání řešitelů projektů „Národní registr VŠKP“ a „Odhalování plagiátů v závěrečných pracích“. Cílem jednání bylo najít společné řešení obou projektů z důvodů jejich obsahové podobnosti, resp. možnost jejich sloučení.

Obě strany se po vyjasnění svých stanovisek a priorit dohodly na podání jednoho společného projektu, který bude obsahovat sjednocení řešení obou dosavadních projektů. Koordinací projektu byla po vzájemné dohodě pověřena Masarykova univerzita, která bude také zodpovídat za technologickou realizaci projektu. Dohodly se též na společném názvu projektu Národní registr vysokoškolských kvalifikačních prací s odhalováním plagiátů.

V kompetenci zástupce Vysoké školy ekonomické v Praze bude návrh uživatelského rozhraní a funkcí národního registru VŠKP, s dodržením potřebných standardů pro oblast workflow, metadat apod. Tato část projektu bude vycházet z doporučení a dosavadních výsledků Odborné komise pro otázky elektronického zpřístupňování vysokoškolských kvalifikačních prací, která bude využívána jako konzultační orgán, jehož doporučeními se zúčastněné školy zavazují řídit. Součástí návrhu národního registru bude také řešení autorskoprávní problematiky, metodických doporučení pro oblast provozu a budování lokálních registrů, citování zdrojů apod.

Pro službu odhalování plagiátů se počítá s využitím systému, který vyvinula a v současné době používá Masarykova univerzita. V rámci projektu bude připraveno otevřené rozhraní pro napojení dalších systémů třetích stran, které by umožnily vyhledávání duplicitních pasáží v plných textech v rámci národního registru. Předpokládá se např. řešení kontroly vůči dalším zdrojům mimo národní registr, kontrola vůči Internetu. Školy samotné budou rozhodovat o tom, který systém kontroly plagiátů využijí, příp. zda vůbec bude kontrola jejich prací prováděna a zda budou poskytovat plné texty prací pro účely kontroly. Projekt počítá jednak se sběrem a zpřístupněním informací o VŠKP, jednak se sběrem

plných textů prací se souhlasem zúčastněné školy za účelem možnosti zjišťovat jejich originalitu. Tato služba bude fungovat na základě dvoustranných smluv mezi danou školou a provozovatelem systému na odhalování plagiátů, příp. provozovatelem národního registru. O způsobu nakládání a zpřístupňování (jak u metadat, tak u plných textů) si bude rozhodovat

každá zúčastněná škola sama. Řešitelé zajistí možnost různých režimů pro vkládání, zpřístupňování plných textů, metadatových údajů o pracích atd.

Vysoké školy, které se dosud hlásily k oběma těmto projektům, se mohou účastnit spolupráce podle svých dosavadních plánů a dílčích projektů. Systém bude do budoucna otevřený i pro všechny další veřejné vysoké školy a bude možné se k němu kdykoliv později bezplatně připojit a využívat jeho služeb (příslib bezplatnosti se netýká nákladů školy na potřebnou úpravu lokálních systémů školy). Předpokládá se, že další rozvoj projektu (např. formou navazujících grantů) bude řešen opět spoluprací současných spoluřešitelů.

Za řešitelský tým MU:	**Za řešitelský tým VŠE a eVŠKP:**
Ing. Michal Brandejs, CSc., MU	Ing. Jan Mach, VŠE
Ing. Jitka Brandejsová, MU	PhDr. Iva Horová, AMU, předsedkyně Komise pro VŠKP

V Brně, 12. září 2007

Příloha II. DART – Europe Dohoda o partnerství

Text dohody podle (41) *přeložil Jan Mach.*

DART – Europe je partnerství výzkumných knihoven a knihovních konsorcií, které společně pracují na zlepšení globálního přístupu k evropských výzkumných pracím. DART – Europe je podporován LIBER (Ligue des Bibliothèques Européennes de Recherche) a je evropskou pracovní skupinou Networked Digital Library of Theses and Dissertations (NDLTD).

Partneři DART – Europe pomáhají poskytovat výzkumným pracovníkům jednotný evropský portál pro vyhledávání elektronických vysokoškolských kvalifikačních prací (eVŠKP) a podílejí se na podpoře ovlivňování budoucího evropského vývoje eVŠKP. DART – Europe nabízí partnerům evropské networkingové fórum k otázkám eVŠKP a může poskytnout příležitost k předložení žádostí o kolaborativní financování k dosažení vize DART – Europe elektronických vysokoškolských kvalifikačních prací.

DART – Europe je financováno z příspěvků partnerů.

Partneři podporují následující principy:

1. DART – Europe bude podporovat vytváření, vyhledávání a využívání evropských eVŠKP a bude udržovat centrální portál pro agregaci a zajištění přístupu k eVŠKP.

2. Evropské knihovny a konsorcia se vyzývají, aby přispěly metadaty do Portálu DART – Europe. Přispěvatelé určí podmínky, za kterých budou metadaty přispívat.

3. DART – Europe vítá od partnerů příspěvek zdroji na podporu správy, vyhledávání, použitelnosti a uchování elektronických prací, a pro prohloubení záměrů a cílů DART – Europe.

4. Partneři jmenují jednoho zástupce, který bude působit jako kontaktní osoba pro DART – Europe, a určí alespoň jednoho zástupce do e-mailové diskuze DART – Europe.

5. DART – Europe vítá nabídky od partnerů hostit příležitostná setkání Projektové rady DART – Europe.

6. Partneři pomáhají zajistit DART – Europe status mezinárodní sítě excelence v oblasti informací, odborných znalostí a zdrojů týkajících se eVŠKP.

7. DART – Europe bude spravováno UCL (University College London) a řízeno Radou složenou ze zástupců partnerských organizací. Stanovy a kompetence správní rady budou stanoveny a čas od času přezkoumávány Radou.

Příloha III. Dotazník Aktuální stav zpřístupňování eVŠKP 2014

Jakým způsobem se zpřístupňují metadata a plné texty VŠKP
na českých veřejných vysokých školách?

Kontakt na realizátora výzkumu a pro dotazy:
Ing. Jan Mach
Vysoká škola ekonomická v Praze, CIKS
tel.: 224 095 586
e-mail: machj@vse.cz

Termín průzkumu:
Prosíme o zaslání vyplněného dotazníku na e-mail machj@vse.cz nejpozději do **29. srpna 2014**.

Pokyny pro vyplnění:
Tento dotazník je určen především správcům repozitářů vysokoškolských kvalifikačních prací. Pokud neznáte odpověď na všechny otázky, prosíme o přeposlání dotazníku odpovědné osobě k doplnění. Dotazník můžete vyplnit částečně, uložit, a až poté přeposlat další osobě k dokončení.
Vyplnění dotazníku zabere přibližně 10-15 minut. Na konci dotazníku je prostor pro upřesnění odpovědí a volný komentář k průzkumu.

Cíl průzkumu:
Průzkum má za cíl poskytnout přehled o vývoji a aktuálním stavu repozitářů vysokoškolských kvalifikačních prací veřejných vysokých škol v ČR. Navazuje na realizované průzkumy Odborné komise pro otázky elektronického zpřístupňování VŠKP AKVŠ z roku 2009 a dílčí dotazníky studentů z let 2010-2013.

Použití výsledků:
Výsledky šetření budou zpracovány na ÚISK FF UK v rámci disertační práce „Správa, vyhledávání a zpřístupňování elektronických vysokoškolských kvalifikačních prací“ a publikovány v odborných fórech a na přednáškách. Poznatky z průzkumu budou využity pro správu a další vývoj repozitářů eVŠKP a navazujících služeb v ČR.

Zkratky a termíny:
- VŠKP – vysokoškolské kvalifikační práce
  (bakalářské, diplomové, rigorózní, disertační a habilitační)
- eVŠKP – vysokoškolské kvalifikační práce v elektronické podobě
- Theses.cz - projekt Národní registr vysokoškolských kvalifikačních prací a systém na odhalování plagiátů
- NUŠL – Národní úložiště šedé literatury
- OAI-PMH – protokol Open Archives Initiative - Protocol for Metadata Harvesting umožňující automatické sklízení metadatových záznamů z repozitářů
- DC – Dublic Core metadatový formát
- MARC - MAchine-Readable Cataloging, rodina standardů pro reprezentaci a komunikaci bibliografických a příbuzných informací ve strojově čitelné formě (MARC 21, MARCXML aj.)

Úvodní informace

1) Jméno a příjmení kontaktní osoby:

2) E-mail kontaktní osoby:
(uveďte prosím kontaktní e-mail na Vás nebo osobu odpovědnou za vyplnění dotazníku)

3) Název vysoké školy: (vyberte Vaši školu)

4) Vámi uváděné odpovědi jsou platné pro:

○ celou vysokou školu

○ pouze pro část školy, prosím upřesněte:

(pokud se odpovědi netýkají celé školy, prosím o přeposlání prázdného dotazníku na další relevantní části školy – děkuji)

5) Jaké typy kvalifikačních prací vybíráte v elektronické podobě?

☐ bakalářské
☐ diplomové
☐ rigorózní
☐ disertační
☐ habilitační
☐ jiné, prosím upřesněte:

6) Platné předpisy školy nebo fakulty stanovují:

☐ způsob zpřístupňování eVŠKP před obhajobou
☐ způsob zpřístupňování posudků
☐ způsob odevzdávání netextových typů prací (umělecká díla, filmy, sochy, …)
☐ workflow (organizace životního cyklu VŠKP a sběru údajů)
☐ formální úpravu prací (úprava titulního listu, …)
☐ požadavky na „technické“ parametry prací (nosiče, formáty, …)
☐ způsob trvalého zpřístupnění po obhajobě
☐ způsob archivace tištěných verzí
☐ způsob archivace elektronických verzí

Pokud kvalifikační práce v elektronické podobě neevidujete ve školním repozitáři ani v theses.cz, zde pro Vás dotazník končí. V případě zájmu můžete doplnit volný komentář v závěrečné otázce dotazníku na poslední stránce.

Sběr a evidence eVŠKP

7) URL repozitáře eVŠKP: http://

8) Který útvar na škole je pověřen řešením problematiky zpřístupňování eVŠKP? Je problematika zpřístupňování řešena centrálně na úrovni univerzity nebo decentralizovaně např. na úrovni fakult?

9) V kterém systému eVŠKP primárně evidujete?

- ○ informační systém školy
- ○ katalog knihovny
- ○ vlastní repozitář školy, specializovaný software (DSpace, Aleph apod.)
- ○ Theses.cz

10) V jaké verzi studenti odevzdávají plný text VŠKP?

- ○ povinně tištěný plný text, elektronická verze je nepovinná
- ○ povinně elektronicky plný text, tištěná verze je nepovinná
- ○ je povinné odevzdat tištěnou i elektronickou verzi plného textu
- ☐ studenti mohou odevzdávat i netextové typy prací (umělecká díla, nahrávky apod.)

11) Jakým způsobem studenti odevzdávají plné texty eVŠKP?

- ☐ odevzdávají soubory na CD/DVD, e-mailem apod. zaměstnanci školy
- ☐ nahrávají přes webové rozhraní (informační systém) školy
- ☐ nahrávají přímo do aplikace Theses.cz
- ☐ jinak:

12) Jakým způsobem je prováděna kontrola eVŠKP na případné projevy plagiátorství?

- ○ kontrolu na projevy plagiátorství neprovádíme
- ○ kontrola se provádí v Theses.cz, výsledky jsou dostupné pouze pověřeným správcům
- ○ kontrola se provádí v Theses.cz, kde mají vyučující výsledky přímo dostupné
- ○ kontrola se provádí v Theses.cz, výsledky jsou dostupné vyučujícím v IS školy
- ○ kontrola se provádí systémem Ephorus
- ○ kontrola se provádí jinak (jiným systémem, kombinací apod.):

Zpřístupňování a export eVŠKP

13) Kdo může získat přístup k plnému textu eVŠKP?

☐ plné texty jsou přístupné studentům/zaměstnancům školy
☐ plné texty jsou přístupné v intranetu školy bez přihlašování
☐ plné texty jsou přístupné veřejnosti po registraci v repozitáři
☐ plné texty jsou přístupné registrovaným čtenářům knihovny
☐ plné texty jsou přístupné registrovaným uživatelům theses.cz/Shibboleth
☐ plné textu jsou volně dostupné na Internetu bez přihlašování aj. omezení
☐ jiné:

14) Na základě jakého oprávnění/licence zpřístupňujete plné texty eVŠKP veřejnosti?

○ plné texty eVŠKP nejsou dostupné veřejnosti
○ na základě licence dané Autorským zákonem (Školní dílo, Knihovní licence apod.)
○ na základě § 47b Zákona o vysokých školách (odevzdáním student dává souhlas)
○ na základě licence Creative Commons, dle volby studenta
○ na základě proprientární licence uzavřené mezi školou a studentem
○ jiné:

15) V případě citlivých nebo utajovaných informací ve VŠKP:

☐ student musí tyto informace vždy anonymizovat a práci zveřejnit celou
☐ jsou tyto informace uvedeny v příloze, která není přístupná veřejnosti
☐ je možné znepřístupnit celou práci
☐ je možné zpřístupnění těchto informací dočasně odložit (časové embargo)
☐ jiné:

16) Jaké mohou být konkrétní důvody pro zvolený způsob ne/zpřístupnění plných textů eVŠKP? Jsou povolené důvody pro nezpřístupnění předem stanoveny, volba je plně na studentovi nebo se řeší individuálně v případě potřeby?

17) Kdo rozhoduje o případném (ne)zpřístupnění práce nebo přílohy veřejnosti?

- ○ student dle vlastního uvážení
- ○ vedoucí práce
- ○ vedoucí katedry
- ○ vedení fakulty (děkan, proděkan)
- ○ vedení školy/rektor
- ○ jiné:

18) Do kterých repozitářů jsou metadata či plné texty eVŠKP průběžně exportovány?

- ☐ katalog knihovny (školy nebo fakulty)
- ☐ Theses.cz
- ☐ NUŠL
- ☐ jiné:

19) Ve kterých formátem jsou metadata eVŠKP z Vašeho repozitáře exportována?
(export mimo školu pro potřeby poskytovatelů služeb – např. Theses.cz, NUŠL, prostřednictvím OAI-PMH serveru aj.)

- ☐ ve standardu Dublin Core (příp. rozšířený o vlastní prvky)
- ☐ ve formátu MARC nebo odvozeném (MARC 21, MARCXML; např. systém Aleph)
- ☐ ve formátu EVSKP-MS
- ☐ ve formátu Theses.cz
- ☐ jiné:

20) Jakým způsobem jsou metadata eVŠKP exportována/předávána mimo školu?
(např. do theses.cz, NUŠL apod.)

- ☐ automatizované sklízení metadat protokolem OAI-PMH poskytovatelem služeb
automatizované odesílání metadat implementované v repozitáři školy podle individuální specifikace třetí strany (např. aplikační rozhraní theses.cz)
- ☐ ad hoc vygenerované soubory předávané třetí straně (ručně, nahráním na web apod.)
- ☐ metadata žádným externím subjektům neposkytujeme
- ☐ jiné:

Děkuji za vyplnění dotazníku

Děkuji Vám za Váš čas věnovaný vyplnění tohoto dotazníkového šetření. Pokud máte jakékoliv doplňující informace nebo připomínky k vyplněnému dotazníku, můžete je uvést zde.

Příloha IV. Mapování prvků EVSKP-MS

LEGENDA K TABULCE

POVINNÉ A OPAKOVATELNÉ PRVKY

znak +
prvky povinné v daném metadatovém formátu

znak *
prvky opakovatelné v daném metadatovém formátu

METADATOVÉ FORMÁTY PRO POPIS eVŠKP

prefix evskp**:** metadatový formát EVSKP-MS

prefix pts**:** metadatový formát Theses.cz

prefix thesis**:** metadatový formát ETD-MS

OBECNÉ METADATOVÉ FORMÁTY

prefix dc**:** nekvalifikovaný Dublin Core Metadata Element Set

prefix dcterms**:** DCMI Metadata Terms

	EVSKP-MS	Theses.cz	ETD-MS	DC Set	DCMI Metadata Terms
Popisné metadatové prvky					
Název VŠKP	dc:title+*	dc:title+ pts:title.translated*	dc:title+*	dc:title*	dc:title*
Podnázev VŠKP	dcterms:alternative*	dcterms:alternative*	dc:title.alternative*	dc:title*	dcterms:alternative*
Autor VŠKP	dc:creator+	dc:creator pts:creator*	dc:creator+*	dc:creator*	dc:creator*
Věcný popis VŠKP	dc:subject*	dc:subject*	dc:subject+*	dc:subject*	dc:subject*
Abstrakt VŠKP	dcterms:abstract+*	dc:description*	dc:description*	dc:description*	dcterms:abstract*
Obsah VŠKP	dcterms:tableOfContents*			dc:description*	dcterms:tableOfContents*
Instituce archivující anebo zpřístupňující VŠKP	dc:publisher*	dc:publisher* pts:publisher.faculcy*	dc:publisher*	dc:publisher*	dc:publisher*
Vedoucí nebo oponent VŠKP	dc:contributor*	pts:advisor* pts:referee*	dc:contributor*	dc:contributor*	dc:contributor*
Datum vytvoření VŠKP	dcterms:created	dcterms:dateSubmitted+	dc:date+	dc:date*	dcterms:created*
Datum odevzdání či podání VŠKP	dcterms:dateSubmitted	dcterms:dateSubmitted+	dc:date+	dc:date*	dcterms:dateSubmitted*
Datum obhajoby VŠKP	dcterms:dateAccepted+	dcterms:dateAccepted	dc:date+	dc:date*	dcterms:dateAccepted*
Datum modifikace VŠKP	dcterms:modified*	pts:getfile.modified		dc:date*	dcterms:modified*
Typ VŠKP	dc:type+*	dc:type+	dc:type*	dc:type*	dc:type*
Médium (formát souboru) VŠKP	dcterms:medium+*	pts:presentation.file-> pts:mtype	dc:format*	dc:format*	dcterms:medium*
Rozsah VŠKP	dcterms:extent*				dcterms:extent*
Identifikátor VŠKP	dc:identifier+*	dc:identifier* pts:thesis.id+	dc:identifier+*	dc:identifier*	dc:identifier*
Jazyk VŠKP	dc:language+*	dc:language+*	dc:language*	dc:language*	dc:language*
Bibliografická citace VŠKP	dcterms: bibliographicCitation*	dcterms: bibliographicCitation*			dcterms: bibliographicCitation*
Práva k využívání VŠKP	dc:rights*	dc:rights* pts:rights.href*	dc:rights*	dc:rights*	dc:rights*

	EVSKP-MS	Theses.cz	ETD-MS	DC Set	DCMI Metadata Terms
Zkratka jména akademického titulu nebo vědecko-pedagogické hodnosti	thesis:degree-> thesis:name+	pts:degree.name+	thesis:degree-> thesis:name*		
Typ studijního programu	thesis:degree-> thesis:level+	pts:degree.level+	thesis:degree-> thesis:level*		
Studijní program a studijní obor	thesis:degree-> thesis:discipline+	pts:degree.discipline+* pts:degree.field*	thesis:degree-> thesis:discipline*		
Instituce přidělující titul	thesis:degree-> thesis:grantor+	pts:degree.grantor+ pts:degree.grantor.faculty *	thesis:degree-> thesis:grantor*		
Technické a administrativní metadatové prvky					
Identifikátor poskytovatele metadat	evskp:contact+	pts:sender.id+			
Počet souborů VŠKP	evskp:fileNumber				
Popis konkrétního souboru VŠKP	evskp:fileProperties*	pts:presentation.file-> pts:ctype pts:mtype pts:size pts:filename pts:fileinfo			
Identifikátor odkazující na soubor tvořící VŠKP nebo archiv ZIP	evskp:transfer*	pts:presentation.file-> pts:url		dc:relation*	dc:relation*
Informace o serveru zpřístupňujícím VŠKP	evskp:server			dc:publisher*	dc:publisher*
Datum doručení metadatového záznamu do repozitáře	evskp:dateDelivered				
Zpřístupnění souborů VŠKP	dcterms:available	pts:presentation.file-> pts:available			dcterms:available
Datum změny záznamu VŠKP	evskp:modified*				dcterms:modified*

Příloha V. OAI-PMH export metadat ve formátu Dublin Core

URL:

http://www.vse.cz/oai/?verb=GetRecord&metadataPrefix=oai_dc&identifier=oai:vse.cz:vskp/4840

<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-
instance" xmlns:dc="http://purl.org/dc/elements/1.1/"xmlns:dcterms="http://purl.org/dc/terms/" xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd http://www.evskp.cz/standardy/evskp/
http://www.evskp.cz/standardy/evskp/1.1/evskp.xsd">
<responseDate>2015-01-21T21:28:14Z</responseDate>
<request verb="GetRecord" metadataPrefix="oai_dc" identifier="oai:vse.cz:vskp/4840">http://www.vse.cz/oai</request>
<GetRecord>
<record>
<header>
<identifier>oai:vse.cz:vskp/4840</identifier>
<datestamp>2014-12-05T05:44:29Z</datestamp>
<setSpec>theses</setSpec>
</header>
<metadata xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:dcterms="http://purl.org/dc/terms/">
<oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:dcterms="http://purl.org/dc/terms/" xmlns:dc="http://purl.org/dc/elements/1.1/"xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
<dc:title xml:lang="cs">Finanční deriváty v zobrazení IFRS</dc:title>
<dc:creator>Remková, Jaroslava</dc:creator>
<dc:description xml:lang="cs">
Cílem této diplomové práce je přiblížit stávající účetní zobrazení finančních derivátů, způsoby jejich oceňování a požadavky kladené na jejich zveřejňování a prezentaci stanovené v rámci Mezinárodních standardů finančního výkaznictví platných k 1. 1. 2007.
</dc:description>
<dc:publisher xml:lang="cs">Vysoká škola ekonomická v Praze</dc:publisher>
<dc:contributor>Vašek, Libor</dc:contributor>
<dc:contributor>Ryneš, Petr</dc:contributor>
<dc:date>2008-02-04</dc:date>
<dc:type xml:lang="eo">info:eu-repo/semantics/masterThesis</dc:type>
<dc:type xml:lang="en">Master's thesis</dc:type>
<dc:type xml:lang="cs">Diplomová práce</dc:type>
<dc:format>application/pdf</dc:format>
<dc:identifier>http://www.vse.cz/vskp/eid/4840</dc:identifier>
<dc:language>cs</dc:language>
<dc:rights xml:lang="cs">
Vysokoškolské kvalifikační práce obhájené na VŠE jsou veřejně dostupné online.
http://ciks.vse.cz/knihovna/Ruzne/vskp_dostupnost.aspx
</dc:rights>
<dc:rights xml:lang="en">
Theses and disertations defended at University of Economics, Prague are freely available online.
http://ciks.vse.cz/knihovna/Ruzne/vskp_dostupnost.aspx
</dc:rights>
</oai_dc:dc>
</metadata>
</record>
</GetRecord>
</OAI-PMH>

Příloha VI. OAI-PMH export metadat ve formátu EVSKP-MS

URL: http://www.vse.cz/oai/
?verb=GetRecord&metadataPrefix=oai_evskpms&identifier=oai:vse.cz:vskp/4840

<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-
instance" xmlns:dc="http://purl.org/dc/elements/1.1/"xmlns:dcterms="http://purl.org/dc/terms/" xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd http://www.evskp.cz/standardy/evskp/
http://www.evskp.cz/standardy/evskp/1.1/evskp.xsd">
<responseDate>2015-01-21T21:25:41Z</responseDate>
<request verb="GetRecord" metadataPrefix="oai_evskpms" identifier="oai:vse.cz:vskp/4840">http://www.vse.cz/oai</request>
<GetRecord>
<record>
<header>
<identifier>oai:vse.cz:vskp/4840</identifier>
<datestamp>2014-12-05T05:44:29Z</datestamp>
<setSpec>theses</setSpec>
</header>
<metadata xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:dcterms="http://purl.org/dc/terms/">
<evskp:metadata xmlns:ccz="http://www.evskp.cz/standardy/corpcz/" xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:dcterms="http://purl.org/dc/terms/"xmlns:dctype="http://purl.org/dc/dcmitype/" xmlns:evskp="http://www.evskp.cz/standardy/evskp/" xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"xmlns:pcz="http://www.evskp.cz/standardy/perscz/" xmlns:thesis="http://www.ndltd.org/standards/metadata/etdms/1.0/" version="1.1">
<dc:title xml:lang="cs">Finanční deriváty v zobrazení IFRS</dc:title>
<dc:creator>Remková, Jaroslava</dc:creator>
<dcterms:abstract xml:lang="cs">
Cílem této diplomové práce je přiblížit stávající účetní zobrazení finančních derivátů, způsoby jejich oceňování a požadavky kladené na jejich zveřejňování a prezentaci stanovené v rámci Mezinárodních standardů finančního výkaznictví platných k 1. 1. 2007.
</dcterms:abstract>
<dc:publisher>
<ccz:universityOrInstitution>
<ccz:name xml:lang="cs">Vysoká škola ekonomická v Praze</ccz:name>
<ccz:email>webmaster@vse.cz</ccz:email>
<ccz:homepage>http://www.vse.cz/</ccz:homepage>
</ccz:universityOrInstitution>
</dc:publisher>
<dc:contributor thesis:role="advisor">Vašek, Libor</dc:contributor>
<dc:contributor thesis:role="referee">Ryneš, Petr</dc:contributor>
<dcterms:created>2007-12-20</dcterms:created>
<dcterms:dateSubmitted>2007-12-20</dcterms:dateSubmitted>
<dcterms:dateAccepted>2008-02-04</dcterms:dateAccepted>
<dcterms:modified>2012-04-30T02:05:00Z</dcterms:modified>
<dc:type xml:lang="cs" evskp:typeType="TypVSKP">Diplomová práce</dc:type>
<dcterms:medium>application/pdf</dcterms:medium>
<dc:identifier>http://www.vse.cz/vskp/eid/4840</dc:identifier>
<dc:language>cs</dc:language>
<thesis:degree>
<thesis:name>Ing.</thesis:name>
<thesis:level xml:lang="cs">Magisterský navazující studijní program</thesis:level>
<thesis:discipline xml:lang="cs">
Finance a účetnictví/Účetnictví a finanční řízení podniku
</thesis:discipline>
<thesis:grantor>
<ccz:universityOrInstitution>
<ccz:name xml:lang="cs">Vysoká škola ekonomická v Praze</ccz:name>
</ccz:universityOrInstitution>

</thesis:grantor>
</thesis:degree>
<evskp:contact contactID="3190"/>
<evskp:fileNumber>3</evskp:fileNumber>
<evskp:fileProperties fileID="1224019" fileType="thesis" fileName="ISIS_7492_xremj02.pdf" fileSize="
657189" format="application/pdf">Hlavní práce</evskp:fileProperties>
<evskp:fileProperties fileID="1224021" fileType="refereeReview" fileName="ISIS_4612_Ryneš.pdf" fil
eSize="674174" format="application/pdf">Oponentura</evskp:fileProperties>
<evskp:fileProperties fileID="1224020" fileType="advisorReview" fileName="ISIS_7492_vasek.pdf" file
Size="85398" format="application/pdf">Hodnocení vedoucího</evskp:fileProperties>
<evskp:transfer accessRights="public" fileID="1224019">
http://www.vse.cz/vskp/show_file.php?soubor_id=1224019
</evskp:transfer>
<evskp:transfer accessRights="public" fileID="1224021">
http://www.vse.cz/vskp/show_file.php?soubor_id=1224021
</evskp:transfer>
<evskp:transfer accessRights="public" fileID="1224020">
http://www.vse.cz/vskp/show_file.php?soubor_id=1224020
</evskp:transfer>
<evskp:server>
<ccz:universityOrInstitution>
<ccz:name xml:lang="cs">Vysoká škola ekonomická v Praze</ccz:name>
<ccz:place xml:lang="cs">Praha</ccz:place>
<ccz:department>
<ccz:name xml:lang="cs">Centrum informačních a knihovnických služeb</ccz:name>
<ccz:email>webmaster@vse.cz</ccz:email>
<ccz:homepage>http://ciks.vse.cz/</ccz:homepage>
</ccz:department>
</ccz:universityOrInstitution>
</evskp:server>
<evskp:modified>2012-04-30T02:05:00Z</evskp:modified>
</evskp:metadata>
</metadata>
</record>
</GetRecord>
</OAI-PMH>

Příloha VII. Soubor schema.xml

Soubor schema.xml vytvořený autorem disertační práce definuje jednotlivá pole indexu kolekce VSKP v Apache Solr, jejich vlastnosti a použité transformace. XML je doplněno o komentáře popisující jednotlivé pasáže.

```
<?xml version="1.0" ?>
<schema name="vskp scheme" version="1.1">
  <!-- types definition -->
  <types>
    <fieldtype name="string"  class="solr.StrField" sortMissingLast="true" omitNorms="true"/>
    <fieldType name="int" class="solr.TrieIntField" precisionStep="0" positionIncrementGap="0"/>
    <fieldType name="long" class="solr.TrieLongField" precisionStep="0" positionIncrementGap="0"/>
    <fieldType name="date" class="solr.TrieDateField" precisionStep="0" positionIncrementGap="0"/>
    <fieldtype name="ignored" stored="false" indexed="false" class="solr.StrField" />

    <!-- values delimited by a semicolon, e.g. dc:subject in EVSKP-MS -->
    <fieldType name="delimitedSemicolon" class="solr.TextField">
        <analyzer>
          <tokenizer class="solr.PatternTokenizerFactory" pattern=";*" />
        </analyzer>
    </fieldType>

    <!–definition of text files in different languages, including stopwords and lematization -->
    <fieldType name="text_general" class="solr.TextField" positionIncrementGap="100">
      <analyzer type="index">
        <tokenizer class="solr.StandardTokenizerFactory"/>
        <filter class="solr.StopFilterFactory" ignoreCase="true" words="stopwords.txt" />
        <!-- in this example, we will only use synonyms at query time
        <filter class="solr.SynonymFilterFactory" synonyms="index_synonyms.txt" ignoreCase="true"
expand="false"/>
        -->
        <filter class="solr.LowerCaseFilterFactory"/>
      </analyzer>
      <analyzer type="query">
        <tokenizer class="solr.StandardTokenizerFactory"/>
        <filter class="solr.StopFilterFactory" ignoreCase="true" words="stopwords.txt" />
        <filter class="solr.StopFilterFactory" ignoreCase="true" words="lang/stopwords_cz.txt" />
        <filter class="solr.SynonymFilterFactory" synonyms="synonyms.txt" ignoreCase="true"
expand="true"/>
        <filter class="solr.LowerCaseFilterFactory"/>
        <filter class="solr.CzechStemFilterFactory"/>
      </analyzer>
    </fieldType>

    <fieldType name="text_en" class="solr.TextField" positionIncrementGap="100">
      <analyzer type="index">
        <tokenizer class="solr.StandardTokenizerFactory"/>
        <!-- in this example, we will only use synonyms at query time
        <filter class="solr.SynonymFilterFactory" synonyms="index_synonyms.txt" ignoreCase="true"
expand="false"/>
        -->
        <!-- Case insensitive stop word removal -->
        <filter class="solr.StopFilterFactory" ignoreCase="true" words="lang/stopwords_en.txt" />
        <filter class="solr.LowerCaseFilterFactory"/>
        <filter class="solr.EnglishPossessiveFilterFactory"/>
        <filter class="solr.KeywordMarkerFilterFactory" protected="protwords.txt"/>
        <filter class="solr.PorterStemFilterFactory"/>
```

```
    </analyzer>
    <analyzer type="query">
      <tokenizer class="solr.StandardTokenizerFactory"/>
      <filter class="solr.SynonymFilterFactory" synonyms="synonyms.txt" ignoreCase="true"
expand="true"/>
      <filter class="solr.StopFilterFactory" ignoreCase="true" words="lang/stopwords_en.txt" />
      <filter class="solr.LowerCaseFilterFactory"/>
          <filter class="solr.EnglishPossessiveFilterFactory"/>
      <filter class="solr.KeywordMarkerFilterFactory" protected="protwords.txt"/>
              <filter class="solr.PorterStemFilterFactory"/>
    </analyzer>
  </fieldType>

  <!-- Czech Tokenizer -->
  <fieldType name="text_cz" class="solr.TextField" positionIncrementGap="100">
    <analyzer>
      <tokenizer class="solr.StandardTokenizerFactory"/>
      <filter class="solr.StopFilterFactory" ignoreCase="true" words="lang/stopwords_cz.txt" />
      <filter class="solr.LowerCaseFilterFactory"/>
      <filter class="solr.CzechStemFilterFactory"/>
      <filter class="solr.StopFilterFactory" ignoreCase="true" words="lang/stopwords_cz.txt" />
    </analyzer>
  </fieldType>

  <!-- other fieldTypes for other languages omitted in this example -->

 </types>

 <!–Definition of extraction rules for EVSK-MS -->
<fields>
 <!-- general -->
 <field name="id" type="string" indexed="true" stored="true" multiValued="false" required="true" />
<field name="_version_" type="long" indexed="true" stored="true" />

<!-- descriptive metadata-->
<field name="title" type="text_general" indexed="true" stored="true" multiValued="true" />
<field name="title_cs" type="text_cz" indexed="true" stored="true" multiValued="true" />
<field name="subject" type="string" indexed="true" stored="true" multiValued="true" />
<field name="language" type="string" indexed="true" stored="true" multiValued="false" />
<field name="type" type="string" indexed="true" stored="true" multiValued="true" />
<field name="abstract" type="text_general" indexed="true" stored="true" multiValued="true" />
<field name="medium" type="string" indexed="true" stored="true" multiValued="false" />

<field name="creator" type="text_general" indexed="true" stored="true" multiValued="true" />
<field name="publisher" type="string" indexed="true" stored="true" multiValued="false" />
<field name="contributor" type="text_general" indexed="true" stored="true" multiValued="true" />
<field name="contact" type="string" indexed="true" stored="true" multiValued="false" />

<!-- Dates -->
<field name="created" type="date" indexed="true" stored="true" multiValued="false" />
<field name="dateSubmitted" type="date" indexed="true" stored="true" multiValued="false" />
<field name="dateAccepted" type="date" indexed="true" stored="true" multiValued="false" />
<field name="modified" type="date" indexed="true" stored="true" multiValued="true" />

<!-- Tag <degree> -->
<field name="degree_name" type="string" indexed="true" stored="true" multiValued="false" />
<field name="degree_level" type="string" indexed="true" stored="true" multiValued="false" />
<field name="degree_discipline" type="string" indexed="true" stored="true" multiValued="false" />
<field name="degree_grantor" type="string" indexed="true" stored="true" multiValued="false" />

<!-- ETD fulltext-->
```

```
 <field name="fulltext" type="text_cz" indexed="true" stored="true" multiValued="true"/> <!-- ETD
fulltext -->
 <dynamicField name="pdf_*" type="string" indexed="true" stored="true" multiValued="true"/> <!-- ETD
fulltext attributes -->

 <!-- XML source information -->
 <field name="file" type="string" indexed="true" stored="true" multiValued="false" />
 <field name="fileAbsolutePath" type="string" indexed="true" stored="true" multiValued="false" />
 <field name="fileLastModified" type="string" indexed="true" stored="true" multiValued="false" />

 <!–Other general variables -->
 <dynamicField name="attr_*" type="text_general" indexed="true" stored="true" multiValued="true"/>
 <dynamicField name="*" type="ignored" multiValued="true" />

 <!-- Prepare fulltext field ‘text’ as a content mix of other fields -->
 <field name="text" type="text_cz" indexed="true" stored="false" multiValued="true" />
 <copyField source="title" dest="text"/>
 <copyField source="title_cs" dest="text"/>
 <copyField source="abstract" dest="text"/>
 <copyField source="subject" dest="text"/>
 <copyField source="fulltext" dest="text"/>
 <copyField source="creator" dest="text"/>
 <copyField source="contributor" dest="text"/>
 </fields>

 <!-- field to use to determine and enforce document uniqueness. -->
 <uniqueKey>id</uniqueKey>

 <!-- field for the QueryParser to use when an explicit fieldname is absent -->
 <defaultSearchField>text</defaultSearchField>

 <!-- SolrQueryParser configuration: defaultOperator -->
 <solrQueryParser defaultOperator="AND"/>
</schema>
```

Příloha VIII. Soubor solrconfig.xml

Soubor solrconfig.xml definuje základní nastavení kolekce VSKP včetně extrakce metadat pomocí obslužné rutiny Data Import Handler.

```
<?xml version="1.0" encoding="UTF-8" ?>
<!--
 Licensed to the Apache Software Foundation (ASF) under one or more
 contributor license agreements.  See the NOTICE file distributed with
 this work for additional information regarding copyright ownership.
 The ASF licenses this file to You under the Apache License, Version 2.0
 (the "License"); you may not use this file except in compliance with
 the License.  You may obtain a copy of the License at

     http://www.apache.org/licenses/LICENSE-2.0

 Unless required by applicable law or agreed to in writing, software
 distributed under the License is distributed on an "AS IS" BASIS,
 WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
 See the License for the specific language governing permissions and
 limitations under the License.
-->

<!--
 This is an example of the rutine to import EVSKP-MS metadata, created by Jan Mach
-->
<config>
  <luceneMatchVersion>4.6</luceneMatchVersion>

  <directoryFactory name="DirectoryFactory"
class="${solr.directoryFactory:solr.StandardDirectoryFactory}"/>

  <dataDir>${solr.vskp.data.dir:}</dataDir>

  <schemaFactory class="ClassicIndexSchemaFactory"/>

  <updateHandler class="solr.DirectUpdateHandler2">
    <updateLog>
      <str name="dir">${solr.vskp.data.dir:}</str>
    </updateLog>
  </updateHandler>

  <requestHandler name="/get" class="solr.RealTimeGetHandler">
    <lst name="defaults">
      <str name="omitHeader">true</str>
    </lst>
  </requestHandler>

  <requestHandler name="/replication" class="solr.ReplicationHandler" startup="lazy" />

  <requestDispatcher handleSelect="true" >
    <requestParsers enableRemoteStreaming="false" multipartUploadLimitInKB="2048"
formdataUploadLimitInKB="2048" />
  </requestDispatcher>

  <requestHandler name="standard" class="solr.StandardRequestHandler" default="true" />
  <requestHandler name="/analysis/field" startup="lazy" class="solr.FieldAnalysisRequestHandler" />
  <requestHandler name="/update" class="solr.UpdateRequestHandler"  />
  <requestHandler name="/admin/" class="org.apache.solr.handler.admin.AdminHandlers" />
```

```
  <requestHandler name="/admin/ping" class="solr.PingRequestHandler">
    <lst name="invariants">
      <str name="q">solrpingquery</str>
    </lst>
    <lst name="defaults">
      <str name="echoParams">all</str>
    </lst>
  </requestHandler>

  <!-- JM: xml files import with a Data Import Handler -->
  <requestHandler name="/dataimport"
class="org.apache.solr.handler.dataimport.DataImportHandler">
      <lst name="defaults">
         <str name="config">dih-config.xml</str>
      </lst>
  </requestHandler>

  <!-- JM: extraction handler -->
  <requestHandler name="/update/extract" class="solr.extraction.ExtractingRequestHandler" >
    <lst name="defaults">
      <str name="fmap.content">text</str>
      <str name="lowernames">true</str>
      <str name="uprefix">attr_</str>
      <str name="captureAttr">true</str>
    </lst>
  </requestHandler>
  <lib dir="lib/extract" regex=".*\.jar" />

  <!-- config for the admin interface -->
  <admin>
    <defaultQuery>solr</defaultQuery>
  </admin>

  <autoCommit>
    <maxDocs>15000</maxDocs>
    <openSearcher>false</openSearcher>
  </autoCommit>

</config>
```

Příloha IX. Soubor dih-config.xml

Soubor dih-config.xml je umístěn v podadresáři conf spravované kolekce VSKP. Obsahuje konfiguraci pro Data Import Request Handler s pravidly extrakce metadat ve formátu EVSKP-MS do jednotlivých polí indexu Apache Solr (viz Příloha VII *Soubor schema.xml*).

```
<dataConfig>
   <dataSource name="dsBinary" type="BinURLDataSource" />
   <dataSource type="FileDataSource" encoding="UTF-8" />

   <script><![CDATA[
    function checkName(row, tag) {
       ret = new java.util.ArrayList();
       aOrig = row.get(tag);

       //index only correct names
       for (var i=0;i < aOrig.size();i++)
       {
         if (aOrig.get(i).indexOf(', ')>0)
         {
           ret.add(aOrig.get(i));
         }
       }

       //in case of count of given names = count of surnames
       var count = row.get('attr_'+ tag +'_surname').size();
       if (row.get('attr_'+ tag +'_forename').size() == count)
       {
         //Add Surname, Given name
         for (var i=0;i < count;i++)
         {
           ret.add(row.get('attr_'+ tag +'_surname').get(i) + ', ' + row.get('attr_'+ tag +'_forename').get(i));
         }
       }

       //swap tag with new values
       row.remove(tag);
       row.put(tag, ret);
       //odstran stare hodnoty
       row.remove('attr_'+ tag +'_forename');
       row.remove('attr_'+ tag +'_surname');

       return row;
    }

    function checkUrl(row)
    {
      var foundUrl= 'http://ciks.vse.cz/download/mach/vskp/empty.pdf';
      //URL for the fulltext

      //if value thesis is in the tag fileProperties
      var pos = row.get('attr_fileProperties_fileType').indexOf('thesis');
      if (pos>-1){
        //get fileID for thesis
        var fileID = row.get('attr_fileProperties_fileID').get(pos);

        //if fileID is found in the tag transfer
        pos = row.get('attr_transfer_fileID').indexOf(fileID);
```

```
        if (pos>-1)
        {
          //set URL to fulltext
          foundUrl = row.get('attr_transfer_url').get(pos);
        }
      }

      row.put('attr_url', foundUrl);

      return row;
    }

    function checkData(row) {
       row = checkName(row, 'creator');
       row = checkName(row, 'contributor');
       row = checkUrl(row);

       return row;
    }
  ]]></script>

<document>
<entity name="files" processor="FileListEntityProcessor" rootEntity="false" dataSource="null"
  fileName="^vse[\w\d-]+\.xml$" baseDir="/opt/solr/cores/vskp/data/xml/" recursive="true"
  newerThan="'NOW-14YEARS'">
  <entity  name="xml" processor="XPathEntityProcessor"
    transformer="RegexTransformer,TemplateTransformer,script:checkData" datasource="files"
    stream="true" forEach="/metadata" useSolrAddSchema="false" xsl="dih-remove-ns.xslt"
    url="${files.fileAbsolutePath}">

  <field column="creator" xpath="/metadata/creator" />
  <field column="attr_creator_forename" xpath="/metadata/creator/person/name/foreName" />
  <field column="attr_creator_surname" xpath="/metadata/creator/person/name/surName" />

  <!-- information about dc:contributor -->
  <field column="contributor" xpath="/metadata/contributor" regex="([\t\n])" replaceWith=""
    flatten="false" />
  <field column="attr_contributor_role" xpath="/metadata/contributor/@role" multiValued="true" />
  <field column="attr_contributor_forename" xpath="/metadata/contributor/person/name/foreName"
    multiValued="true" />
  <field column="attr_contributor_surname" xpath="/metadata/contributor/person/name/surName"
    multiValued="true" />

   <field column="id" xpath="/metadata/identifier" regex="(\d+)" />
  <field column="title_cs" xpath="/metadata/title" />
  <field column="subject" xpath="/metadata/subject" splitBy="; " />
  <field column="abstract" xpath="/metadata/abstract" />
  <field column="created" xpath="/metadata/created" template="${xml.created}T00:00:00Z" />
  <field column="dateSubmitted" xpath="/metadata/dateSubmitted"
    template="${xml.dateSubmitted}T00:00:00Z" />
  <field column="modified" xpath="/metadata/modified"
    regex="(\d{4})-(\d{2})-(\d{2}) (\d{2}):(\d{2}):(\d{2})" />

  <!-- degree -->
  <field column="degree_name" xpath="/metadata/degree/name" />
  <field column="degree_level" xpath="/metadata/degree/level" />
  <field column="degree_discipline" xpath="/metadata/degree/discipline" />
  <field column="degree_grantor" xpath="/metadata/degree/grantor" />
```

```
    <!-- fileProperties -->
    <field column="attr_fileProperties_fileID" xpath="/metadata/fileProperties/@fileID"
      multiValued="true" />
    <field column="attr_fileProperties_fileType" xpath="/metadata/fileProperties/@fileType"
      multiValued="true" />
    <field column="attr_fileProperties_format" xpath="/metadata/fileProperties/@fileType"
      multiValued="true" />
    <field column="attr_fileProperties_value" xpath="/metadata/fileProperties" multiValued="true" />

    <!-- transfer -->
    <field column="attr_transfer_fileID" xpath="/metadata/transfer/@fileID" multiValued="true" />
    <field column="attr_transfer_url" xpath="/metadata/transfer" multiValued="true" />

    <!-- fulltext extraction -->
    <entity processor="TikaEntityProcessor" name="tika" format="text"
      url="http://solr.vse.cz/download/mach/vskp/proxy.ashx?url=${xml.attr_url}"
      dataSource="dsBinary" onError="skip" transformer="RegexTransformer">
      <field name="pdf_contentType" column="Content-Type" meta="true" />
      <field name="pdf_pages" column="xmpTPg:NPages" meta="true" />
      <field name="fulltext" column="text" regex="\n" replaceWith=" "/>
    </entity>

    <!-- XML extraction -->
    <entity processor="PlainTextEntityProcessor" name="txt" url="${files.fileAbsolutePath}"
      dataSource="files">
      <field column="plainText" name="attr_plaintext1"/>
     </entity>
   </entity>
  </entity>
 </document>
</dataConfig>
```

Příloha X. Soubor dih-remove-ns.xslt

Při importu XML ve formátu EVSKP-MS do Apache Solr bylo zapotřebí použít XSLT transformace připravené autorem podle (113), jejímž cílem je odstranění použitých jmenných prostorů ve formátu EVSKP-MS. Níže uvedená XSLT transformace je uložena v souboru dih-remove-ns.xslt v adresáři conf kolekce.

```
<?xml version="1.0"?>
<xsl:stylesheet version="1.0" xmlns:xsl="http://www.w3.org/1999/XSL/Transform">

   <xsl:output indent="yes" method="xml" encoding="utf-8" omit-xml-declaration="yes"/>

   <!-- kopírování elementů -->
   <xsl:template match="*">
      <xsl:element name="{local-name()}">
         <xsl:apply-templates select="@* | node()"/>
      </xsl:element>
   </xsl:template>

   <!-- kopírování atributů -->
   <xsl:template match="@*">
      <xsl:attribute name="{local-name()}">
         <xsl:value-of select="."/>
      </xsl:attribute>
   </xsl:template>
</xsl:stylesheet>
```

Příloha XI. Kompletní výsledky testů systémů na detekci duplicit

Vyhodnocení testů detekce plagiátorství

Metadata

		Zdroj										Keywords					Format			Language			Created		
Č.	ID	W-CZ	W-EN	T-CZ	T-EN	NDLTD	ANO	ARX	G-CZ	G-EN	EIZ	LTE	UMTS	EDGE	WIMAX	Wi-Fi	PDF	html	doc	cz	en	sk	2013	2012	other
1	a1	+										+						+		+			+		
2	a2	+											+					+		+			+		
3	a3	+												+				+		+			+		
4	a4	+													+			+		+			+		
5	a5	+														+		+		+			+		
11	b1		+									+						+			+		+		
12	b2		+										+					+			+		+		
13	b3		+											+				+			+		+		
14	b4		+												+			+			+		+		
15	b5		+													+		+			+		+		
21	c1			+								+					+			+					+
22	c2			+									+				+			+				+	
23	c3			+										+			+			+					+
24	c4			+											+		+			+				+	
25	c5			+												+	+			+				+	
31	d1				+							+					+				+				+
32	d2				+								+				+				+		+		
33	d3				+									+					+		+				+
34	d4				+										+		+				+				+
35	d5				+											+	+				+				+
41	e1					+						+					+				+				+
42	e2					+							+				+				+				+
43	e3					+								+			+				+			+	
44	e4					+									+		+				+				+
45	e5					+										+	+				+				+
51	f1						+					+						+		+				+	
52	f2						+						+					+		+				+	
53	f3						+							+				+		+			+		
54	f4						+								+			+				+		+	
55	f5						+									+		+		+			+		
61	g1							+				+					+				+		+		
62	g2							+					+				+				+		+		
63	g3							+						+			+				+		+		
64	g4							+							+		+				+			+	
65	g5							+								+	+				+			+	
71	h1								+			+						+		+			+		
72	h2								+				+					+		+					+
73	h3								+					+				+		+			+		
74	h4								+						+			+		+			+		
75	h5								+							+		+		+			+		
81	i1									+		+						+			+		+		
82	i2									+			+					+			+			+	
83	i3									+				+				+			+		+		
84	i4									+					+			+			+		+		
85	i5									+						+		+			+		+		
91	j1										+	+					+				+				+
92	j2										+		+				+				+				+
93	j3										+			+			+				+				+
94	j4										+				+		+				+				+
95	j5										+					+	+				+				+
∑	+	5	5	5	5	5	5	5	5	5	5	10	10	10	10	10	24	25	1	19	30	1	24	10	16
	o																								

Zdroje

		THESES.CZ								Turnitin								EPHORUS								GooglePlagiarism							
Č.	ID	Věta	Odstavec	Prohození	Diakritika	Par. věty	Par. slova	Transl.	SUMA	Věta	Odstavec	Prohození	Diakritika	Par. věty	Par. slova	Transl.	SUMA	Věta	Odstavec	Prohození	Diakritika	Par. věty	Par. slova	Transl.	SUMA	Věta	Odstavec	Prohození	Diakritika	Par. věty	Par. slova	Transl.	SUMA
1	a1									+	+	+	+		+		+		o		o				+	+	o						+
2	a2	+	+	+	+		o		+	+	+	+	+		+		+										+						+
3	a3	+	+	o	+		+		+	+	+	+	+		+		+									+	o		+				+
4	a4									+	+	+	+		+		+							o	+	+	o		+				+
5	a5		+						+	o	+	o	o		o		+									+	o		+				+
11	b1	+	+	o			+		+	o	+	o		+	o		+		o	o					+	+	o				+		+
12	b2									o	o	o		o	o		+									+	o						+
13	b3									o	o	o		o	o		+									+	o						+
14	b4																		+				o		+	+	+						+
15	b5																									+	+						+
21	c1	o	o	o	o				+	o	o	o	o		o		+	o							+	o	o		o				+
22	c2										o						+																
23	c3																																
24	c4																																
25	c5																																
31	d1									o	o	o		o	o		+																
32	d2									o	o	o		o	o		+																
33	d3																									+	+						+
34	d4																																
35	d5									+	+	+			+		+									+							+
41	e1																									+							+
42	e2																																
43	e3																																
44	e4																																
45	e5																																
51	f1																																
52	f2																																
53	f3																																
54	f4																																
55	f5																																
61	g1																																
62	g2																									+							+
63	g3																																
64	g4									o	o	o		o	o		+									+							+
65	g5																									o							+
71	h1	+	+	o	+		+		+																	+	+		+				+
72	h2									+	+	+	+		+		+									+	+		+				+
73	h3																									+	o						+
74	h4	+	+	o	+				+	+	+	o	+		+		+									+	+		+				+
75	h5									o	o	o	o		o		+									+	+		+				+
81	i1									+	+	+		+	+		+									+	+						+
82	i2									o		o		o	o		+									+	o						+
83	i3																									+	o						+
84	i4																																
85	i5																																
91	j1									o	o	o		o	o		+																
92	j2									o	o	o		o	o		+									+	o						+
93	j3										o						+									o							+
94	j4																									+	o						+
95	j5																									+	o						+
∑	+	5	6	1	4	0	3	0	7	8	10	7	6	2	8	0	22	0	1	0	0	0	0	0	5	25	9	0	7	0	1	0	29
	o	1	1	5	1	0	1	0	0	12	11	13	3	8	12	0	0	1	2	1	1	0	1	1	0	3	14	0	1	0	0	0	0

Tabulka 24 Kompletní výsledky testů systémů na detekci duplicit (zdroj: vlastní zpracování)

Kompletní tabulka s výsledky provedených testů systémů na detekci duplicit je dostupná online:

http://www.researchgate.net/publication/273144181

DOI: 10.13140/2.1.4398.4162

Příloha XII. Zdroje použité v testu systémů na detekci duplicit

a1) LTE. In: *Wikipedie: otevřená encyklopedie* [online]. San Francisco (CA): Wikimedia Foundation, 5. 4. 2013 [cit. 18. 4. 2013]. Dostupné z: http://cs.wikipedia.org/wiki/LTE

a2) Universal Mobile Telecommunications System. In: *Wikipedie: otevřená encyklopedie* [online]. San Francisco (CA): Wikimedia Foundation, 7. 4. 2013 [cit. 18. 4. 2013]. Dostupné z: http://cs.wikipedia.org/wiki/UMTS

a3) Enhanced Data Rates for GSM Evolution. In: *Wikipedie: otevřená encyklopedie* [online]. San Francisco (CA): Wikimedia Foundation, 5. 4. 2013 [cit. 18. 4. 2013]. Dostupné z: http://cs.wikipedia.org/wiki/Enhanced_Data_Rates_for_GSM_Evolution

a4) WiMAX. In: *Wikipedie: otevřená encyklopedie* [online]. San Francisco (CA): Wikimedia Foundation, 8. 3. 2013 [cit. 18. 4. 2013]. Dostupné z: http://cs.wikipedia.org/wiki/WiMAX

a5) Wi-Fi. In: *Wikipedie: Otevřená encyklopedie* [online]. San Francisco (CA): Wikimedia Foundation, 16. 4. 2013 [cit. 18. 4. 2013]. Dostupné z: http://cs.wikipedia.org/wiki/Wi-fi

b1) LTE (telecommunication). In: *Wikipedia: the free encyclopedia* [online]. San Francisco (CA): Wikimedia Foundation, 15. 4. 2013 [cit. 18. 4. 2013]. Dostupné z: http://en.wikipedia.org/wiki/LTE_(telecommunication)

b2) Universal Mobile Telecommunications System. In: *Wikipedia: the free encyclopedia* [online]. San Francisco (CA): Wikimedia Foundation, 15. 4. 2013 [cit. 18. 4. 2013]. Dostupné z: http://en.wikipedia.org/wiki/Umts

b3) Enhanced Data Rates for GSM Evolution. In: *Wikipedia: the free encyclopedia* [online]. San Francisco (CA): Wikimedia Foundation, 6. 3. 2013 [cit. 18. 4. 2013]. Dostupné z: http://en.wikipedia.org/wiki/Enhanced_Data_Rates_for_GSM_Evolution

b4) WiMAX. In: *Wikipedia: the free encyclopedia* [online]. San Francisco (CA): Wikimedia Foundation, 10. 4. 2013 [cit. 18. 4. 2013]. Dostupné z: http://en.wikipedia.org/wiki/WiMAX

b5) Wi-Fi. In: *Wikipedia: the free encyclopedia* [online]. San Francisco (CA): Wikimedia Foundation, 28. 5. 2013 [cit. 1. 6. 2013]. Dostupné z: http://en.wikipedia.org/wiki/Wi-fi

c1) HAVLÍČKOVÁ, Klára. *Analýza současných služeb vybraného mobilního operátora* [online]. Diplomová práce. Vysoká škola ekonomická v Praze. 31. 10. 2011 [cit. 24. 4. 2013]. Dostupné z: http://isis.vse.cz/zp/113975

c2) BURIAN, Martin. *Marketingová analýza produktu mobilní data společnosti T-Mobile ČR* [online]. Bakalářská práce. Vysoká škola ekonomická v Praze. 15. 5. 2012 [cit. 24. 4. 2013]. Dostupné z: http://isis.vse.cz/zp/91837

c3) KOPŘIVA, Ondřej. *Edge of Darkness - analýza seriálu* [online]. Bakalářská práce. Univerzita Palackého v Olomouci. 4. 5. 2011 [cit. 24. 4. 2013]. Dostupné z: http://Theses.cz/id/klcev6/Edge_of_darkness.pdf

c4) VÝBORNÝ, Vojtěch. *Mobilní internet v ČR a ve světě* [online]. Bakalářská práce. Vysoká škola ekonomická v Praze. 18. 1. 2012 [cit. 1. 6. 2013]. Dostupné z: http://isis.vse.cz/zp/92924

c5) MATLAS, Jiří. *Návrh a výstavba moderní datové Wi-Fi sítě pro ISP* [online]. Diplomová práce. Jihočeská univerzita v Českých Budějovicích. 27. 4. 2012 [cit. 24. 4. 2013]. Dostupné z: http://Theses.cz/id/yd7s2x/Matlas_Ji_Nvrh_a_vstavba_modern_datov_Wi-Fi_st_pro_ISP.pdf

d1) STRÁŽNICKÝ, Matúš. *Application of Game Theory principles in the oligopoly-characterized industry* [online]. Diplomová práce. Vysoká škola ekonomická v Praze. 23. 5. 2011 [cit. 24. 4. 2013]. Dostupné z: http://isis.vse.cz/zp/116876

d2) KOHUT, Michal. *OpenIMS Modelling for Performance Analysis* [online]. Diplomová práce. Vysoká škola ekonomická v Praze. 8. 2. 2013 [cit. 24. 4. 2013]. Dostupné z: http://is.muni.cz/th/207951/fi_m/Thesis.pdf

d3) URBANCOVÁ, Žaneta. *Translation and Analyses of Bridget Jones: The Edge of Reason by Helen Fielding* [online]. Diplomová práce. Masarykova univerzita. 14. 6. 2011 [cit. 24. 4. 2013]. Dostupné z: http://is.muni.cz/th/104208/pedf_m/Diploma_Theses.docx

d4) CHOCHOLOVÁ, Petra. *Trends in Mobile Marketing* [online]. Diplomová práce. Vysoká škola ekonomická v Praze. 1. 1. 1990 [cit. 3. 6. 2011]. Dostupné z: http://isis.vse.cz/zp/94341

d5) ŘEHŮŘEK, Radim. *Scalability of Semantic Analysis in Natural Language Processing* [online]. Disertační práce. Masarykova univerzita. 12. 9. 2011 [cit. 24. 4. 2013]. Dostupné z: http://is.muni.cz/th/39672/fi_d/dizertace.pdf

e1) SALAH, Mohamed. *Comparative Performance Study of LTE Uplink Schedulers* [online]. Diplomová práce. Queen's University. 7. 5. 2011 [cit. 24. 4. 2013]. Dostupné z: http://qspace.library.queensu.ca/bitstream/1974/6509/3/SALAH_Mohamed_201104_MASc.pdf

e2) PILLAI, Anju. *A Connection Admission Control Framework for UMTS based Satellite Systems* [online]. Disertační práce. University of Bradford. 1. 1. 2011 [cit. 24. 4. 2013]. Dostupné z: http://bradscholars.brad.ac.uk/bitstream/handle/10454/5487/Thesis_APillai.pdf?sequence=1

e3) MCGOOGAN, Keriann C. *Edge Effects on the Behaviour and Ecology of Propithecus coquereli in Northwest Madagascar* [online]. Disertační práce. University of Toronto. 10. 1. 2012 [cit. 24. 4. 2013]. Dostupné z: http://hdl.handle.net/1807/31861

e4) LIU, Bin. *Mécanismes de handover inter système 3G-WiMAX. Etude des performances comparées d'une approche basée sur IP et d'une approche utilisant des protocoles radio de niveau 2* [online]. Disertační práce. Paristech. 4. 5. 2009 [cit. 24. 4. 2013]. Dostupné z: http://pastel.archives-ouvertes.fr/docs/00/50/13/91/PDF/Thesis_BinLIU_new.pdf

e5) DRISKELL, Luke. *Mapping the digital divide in neighborhoods: Wi-Fi access in Baton Rouge, Louisiana* [online]. Master's Thesis. Louisiana State University. 26. 3. 2010 [cit. 24. 4. 2013]. Dostupné z: http://etd.lsu.edu/docs/available/etd-04282010-101710/unrestricted/driskell_thesis.pdf

f1) Další krok v evoluci mobilní komunikace. In: *Technik* [online]. 8. 1. 2012 [cit. 26. 4. 2013]. Dostupné prostřednictvím Anopress IT: http://www.anopress.cz/Web/Gate/artc.aspx?idfk=COTN20121108010017&hash=ea6nraxfys&lang=cz

f2) Sítě třetí generace. In: *Computerworld* [online]. 18. 5. 2012 [cit. 26. 4. 2013]. Dostupné prostřednictvím Anopress IT: http://www.anopress.cz/Web/Gate/artc.aspx?idfk=CTCW20120518010027&hash=pxqt6pytqv&lang=cz

f3) ECKSTEIN, Michael. Nejlepší TV do obývacího pokoje. In: *Chip* [online]. 18. 1. 2013 [cit. 26. 4. 2013]. Dostupné prostřednictvím Anopress IT: http://www.anopress.cz/Web/Gate/artc.aspx?idfk=CTCH20130118020059&hash=nrrxb8bmlj&lang=cz

f4) LACKO, Ľuboslav. Bezdrôtová komunikácia a siete. In: *Reseller Magazine* [online]. 5. 12. 2012 [cit. 26. 4. 2013]. Dostupné prostřednictvím Anopress IT: http://www.anopress.cz/Web/Gate/artc.aspx?idfk=CTRM20121205010013&hash=kkcsjj7wli&lang=cz

f5) DORŇÁK, Radek. Na cesty výhodně. In: *Computer* [online]. 14. 2. 2013 [cit. 26. 4. 2013]. Dostupné prostřednictvím Anopress IT: http://www.anopress.cz/Web/Gate/artc.aspx?idfk=CTC420130214010004&hash=phkyy06j5t&lang=cz

g1) PARRUCA, *Donald aj. Analytical Model of Proportional Fair Scheduling in Interference-limited OFDMA/LTE Networks* [online]. 7. 3. 2013 [cit. 26. 4. 2013]. Dostupné z: http://arxiv.org/pdf/1303.1778.pdf

g2) TSAY, Joe-Kai a Stig MJØLSNES. *Computational Security Analysis of the UMTS and LTE Authentication and Key Agreement Protocols* [online]. 8. 1. 2013 [cit. 26. 4. 2013]. Dostupné z: http://arxiv.org/pdf/1203.3866.pdf

g3) MARX, Dániel a László A. VÉGH. *Fixed-parameter algorithms for minimum cost edge-connectivity augmentation* [online]. 24. 4. 2013 [cit. 26. 4. 2013]. Dostupné z: http://arxiv.org/pdf/1304.6593.pdf

g4) ABDALLAH, Ayman aj. *Design and Performance Study of Smart Antenna Systems for WIMAX Applications* [online]. 25. 12. 2012 [cit. 26. 4. 2013]. Dostupné z: http://arxiv.org/ftp/arxiv/papers/1212/1212.6056.pdf

g5) KANG, Du Ho aj. *Cost Efficient High Capacity Indoor Wireless Access: Denser Wi-Fi or Coordinated Pico-cellular?* [online]. 19. 11. 2012 [cit. 26. 4. 2013]. Dostupné z: http://arxiv.org/pdf/1211.4392.pdf

h1) T-Mobile – LTE. *T-Mobile* [online]. 26. 4. 2013 [cit. 26. 4. 2013]. Dostupné z: http://www.t-mobile.cz/web/cz/osobni/internet/nejrychlejsi-internet/lte

h2) Jak se vyznat v mobilních datových sítích (UMTS, HSDPA, HSUPA, HSPA+, LTE). NITANA. *BusinessVize* [online]. 29. 6. 2010 [cit. 26. 4. 2013]. Dostupné z: http://www.businessvize.cz/datove-prenosy-a-site/jak-se-vyznat-v-mobilnich-datovych-sitich-umts-hsdpa-hsupa-hspa-lte

h3) Will the Circle…. *Living at the edge of the world* [online]. 9. 4. 2013 [cit. 26. 4. 2013]. Dostupné z: http://www.edgeoftheworld.cz/2013/04/09/will-the-circle/

h4) Hlavní stránka. *Wimax Networking* [online]. 29. 1. 2013 [cit. 26. 4. 2013]. Dostupné z: http://www.wimax.cz/

h5) WIFI Czech Republic. *WIFI Czech Republic* [online]. 26. 4. 2013 [cit. 26. 4. 2013]. Dostupné z: http://www.wifi-cz.cz/CZ/cz/O nás/WIFI Czech Republic/WIFI_CZ_cz.aspx

i1) LTE. *3GPP* [online]. 26. 4. 2013 [cit. 26. 4. 2013]. Dostupné z: http://www.3gpp.org/LTE

i2) HALLIWELL, Donny. HSPA, UMTS, GSM, LTE and Other Acronyms Demystified. *Inside Blackbarry* [online]. 21. 8. 2012 [cit. 26. 4. 2013]. Dostupné z: http://blogs.blackberry.com/2012/08/smartphone-acronyms/

i3) EDGE. *Edge* [online]. 25. 4. 2013 [cit. 26. 4. 2013]. Dostupné z: http://www.edge.org/

i4) Home. *WiMax.com: 4G Wireless Broadband Solutions* [online]. 26. 4. 2013 [cit. 26. 4. 2013]. Dostupné z: http://www.wimax.com/

i5) Discover and Learn. *WI-FI Alliance* [online]. 1. 1. 2013 [cit. 1. 6. 2013]. Dostupné z: http://www.wi-fi.org/discover-and-learn

j1) KANG, Ting-Wei aj. Internal mobile phone antenna array for LTE/WWAN and LTE MIMO operations. In: *Antennas and Propagation (EUCAP), Proceedings of the 5th European Conference on on Antennas and Propagation* [online]. 1. 7. 2011 [cit. 31. 5. 2013]. Dostupné z: http://web.ebscohost.com/ehost/detail?bdata=Jmxhbmc9Y3Mmc2l0ZT1laG9zdC1saXZl#db=a9h&AN=60135404

j2) EMMENEGGER, Mireille Faist. *Life Cycle Assessment of the Mobile Communication System UMTS: Towards Eco-efficient Systems* [online]. 1. 7. 2006 [cit. 31. 5. 2013]. Dostupné z: http://search.proquest.com/docview/664755681/13E60E4DD4316DF6253/1?accountid=17203

j3) CHRISTIAN, Jeffrey E. *The Headhunter's Edge* [online]. 1. 9. 2002 [cit. 31. 5. 2013]. Dostupné z: http://site.ebrary.com/lib/vsep/Doc?id=10021870&ppg=15

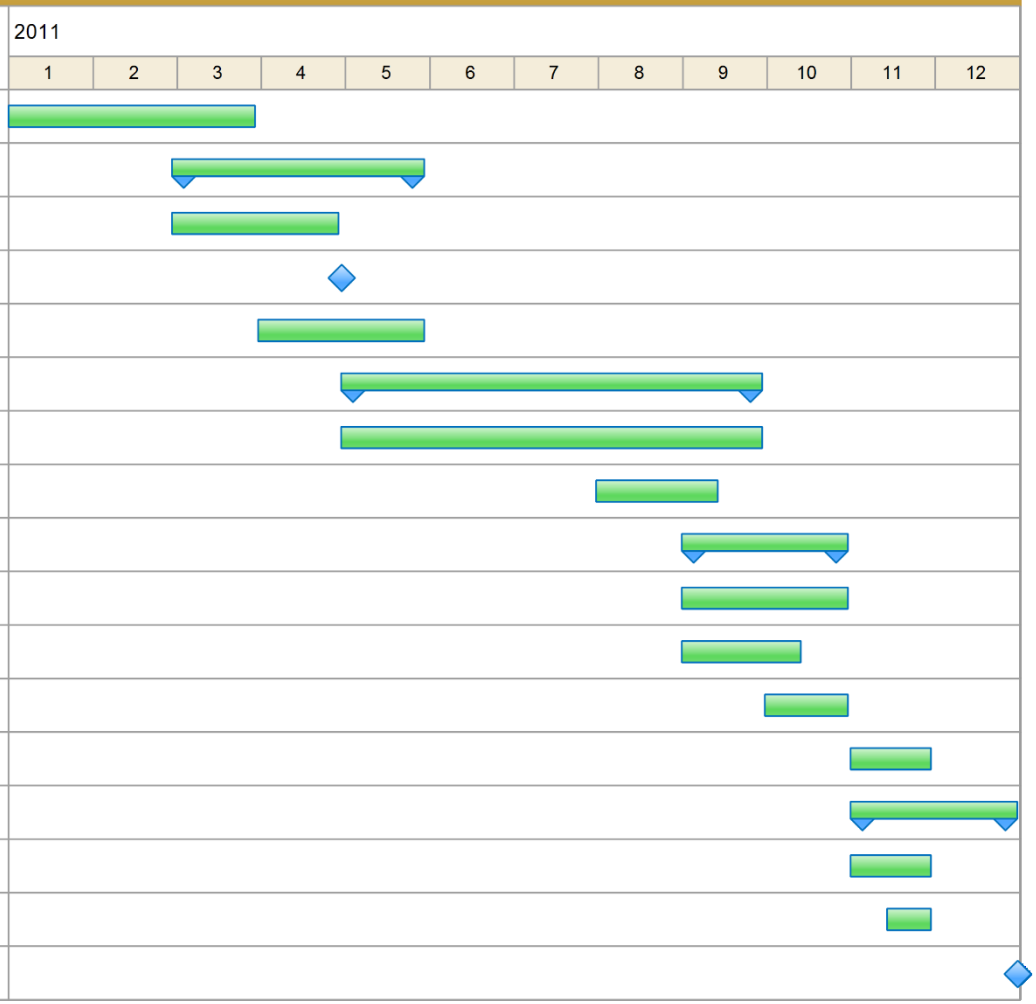
j4) WELLMAN, Barry. The Reconstruction of Space and Time: Mobile Communication Practices. *Contemporary Sociology* [online]. 1. 3. 2010 [cit. 31. 5. 2013]. Dostupné z: http://www.jstor.org/stable/20695348

j5) Development of Voice over WiFi by Integrating Mobile Networks. OECD digital economy papers. In: *OECD iLibrary* [online]. 14. 4. 2005 [cit. 31. 5. 2013]. Dostupné z: http://www.oecd-ilibrary.org/docserver/download/5kz9j9bnn020.pdf?expires=1370019896&id=id&accname=guest&checksum=2A725230E6D5C49FAA1451776B045EC7

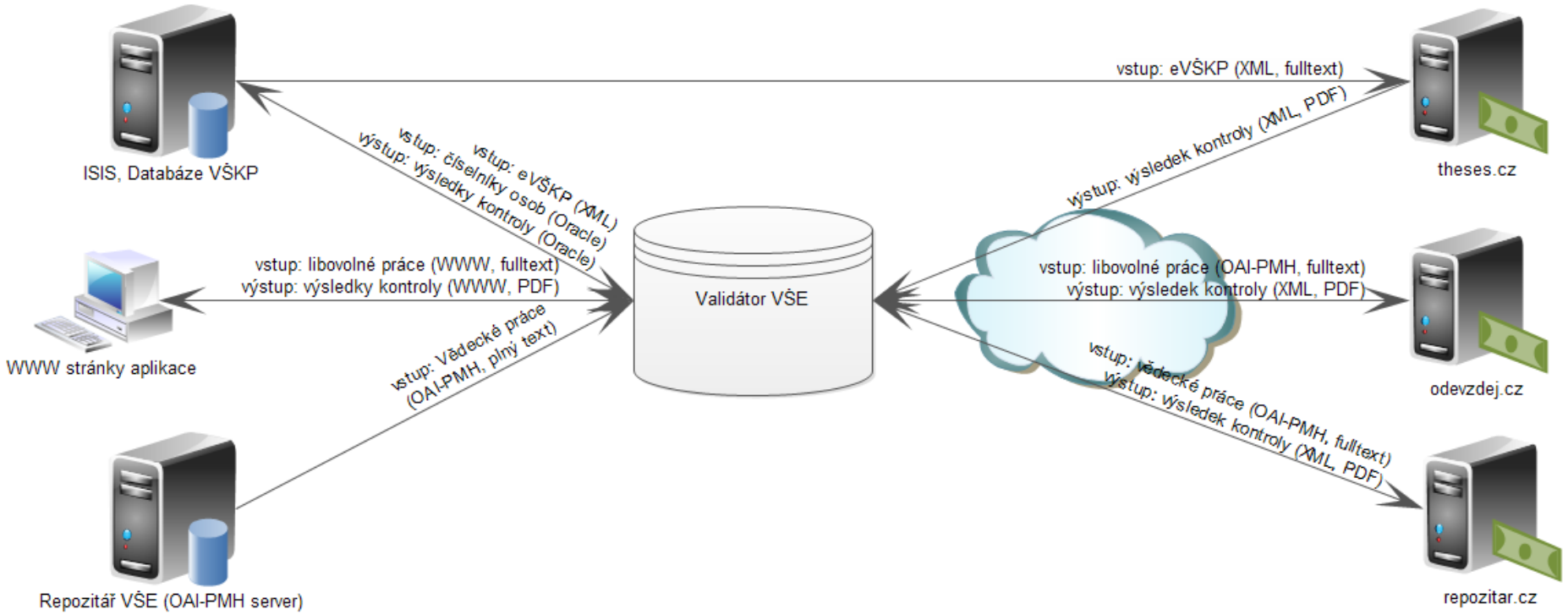
Příloha XIII. Ganttův diagram projektu Validátor VŠE

Jan 1 2011 - Dec 31 2011 — Validátor VŠE

No.	Task Name	Start	Finish	Duration
1	Příprava projektu	1. 1. 2011	31. 3. 2011	89.0 d.
2	Zadání projektu	1. 3. 2011	31. 5. 2011	91.0 d.
3	Výběr dodavatele	1. 3. 2011	30. 4. 2011	60.0 d.
4	Smlouva	1. 5. 2011	1. 5. 2011	0.0 d.
5	Zadávací dokumentace	1. 4. 2011	31. 5. 2011	60.0 d.
6	Příprava SW a HW	1. 5. 2011	30. 9. 2011	152.0 d.
7	Programování	1. 5. 2011	30. 9. 2011	152.0 d.
8	Posílení HW infrastruktury	1. 8. 2011	14. 9. 2011	44.0 d.
9	Ověřovací provoz	1. 9. 2011	31. 10. 2011	60.0 d.
10	Testování aplikace	1. 9. 2011	31. 10. 2011	60.0 d.
11	Příprava metod. opatření	1. 9. 2011	10/14/2011	43.0 d.
12	Předání projektu	1. 10. 2011	31. 10. 2011	30.0 d.
13	Vyúčtování projektu	1. 11. 2011	30. 11. 2011	29.0 d.
14	Ukončení projektu	1. 11. 2011	31. 12. 2011	60.0 d.
15	Schválení metod. opatření	1. 11. 2011	30. 11. 2011	29.0 d.
16	Reportování projektu	14. 11. 2011	30. 11. 2011	16.0 d.
17	Předání aplikace do provozu	31. 12. 2011	31. 12. 2011	0.0 d.

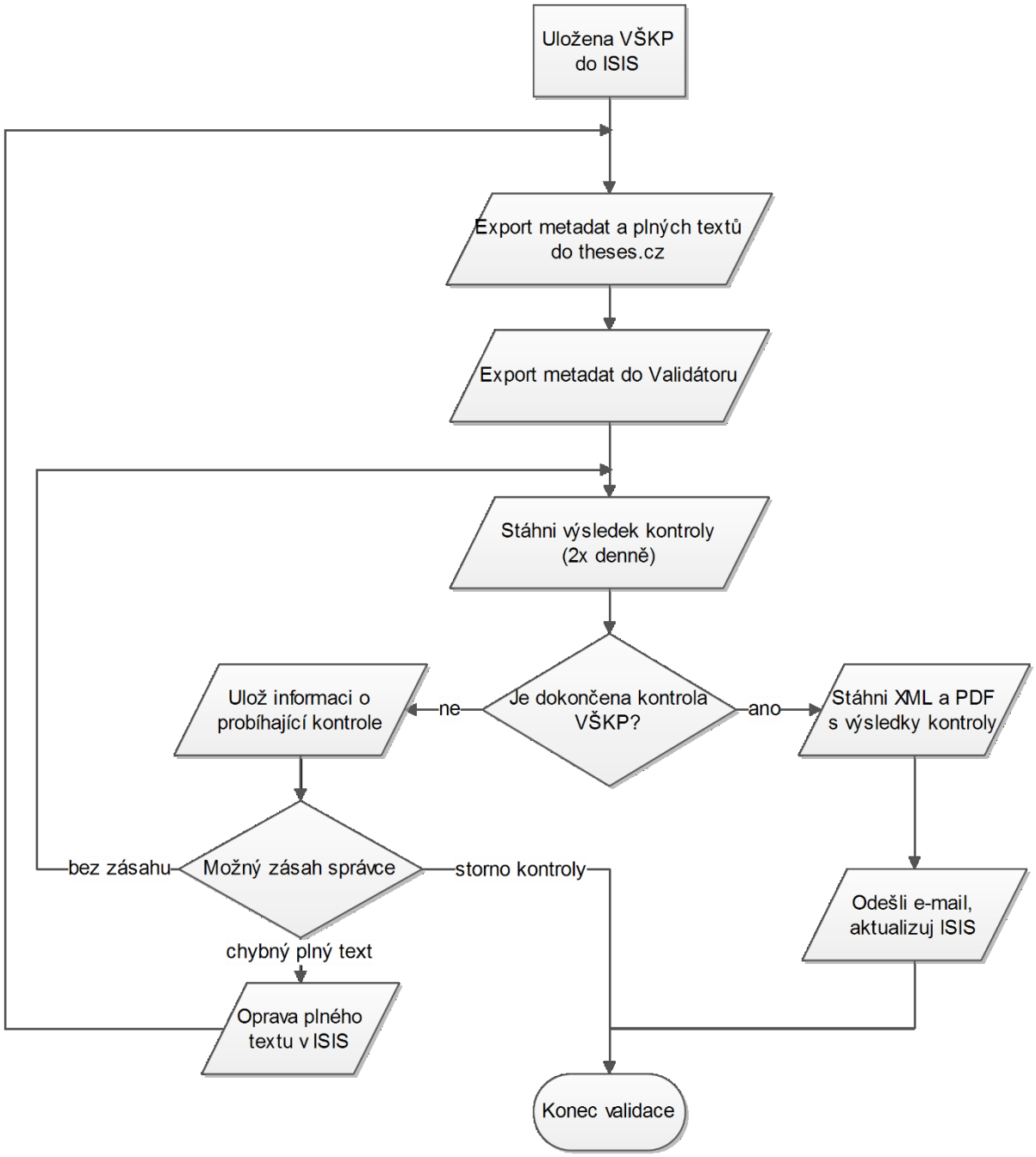


Příloha XIV. Síťový graf vstupů a výstupů aplikace Validátor VŠE



Obrázek 29 Síťový graf vstupů a výstupů aplikace Validátor VŠE (zdroj: autor)

Příloha XV. Vývojový diagram kontroly eVŠKP na VŠE v Praze



Obrázek 30 Vývojový diagram kontroly eVŠKP na VŠE v Praze (zdroj: autor)