Univerzita Karlova v Praze

Filozofická fakulta

Ústav informačních studií a knihovnictví

Diplomová práce

Bc. Michaela Charvátová

Sjednocování věcného popisu agregovaných záznamů v repozitáři NUŠL

Unification of Subject Description of Aggregated Records in National Repository of Grey Literature

Praha 2016

Vedoucí práce: PhDr. Eva Bratková, Ph.D.

Prohlašuji, že jsem diplomovou práci vypracovala samostatně, že jsem řádně citovala 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

Klíčová slova (česky): digitální repozitáře, věcný popis, mapování, automatická indexace, Národní úložiště šedé literatury

Abstrakt (česky):

Diplomová práce se zabývá metodami sjednocení věcného popisu v záznamech agregovaných z různých zdrojů v prostředí digitálního repozitáře na příkladu Národního úložiště šedé literatury (NUŠL). Po představení zahraničních zkušeností ze systémů BASE a LASSO je popsána i současná praxe v repozitáři NUŠL, v němž je k jednotnému popisu pomocí Polytematického strukturovaného hesláře (PSH) využívána automatická indexace.

V rámci práce byly na PSH namapovány skupiny Konspektu a tezaurus MeSH. Tato mapování byla aplikována na záznamy přebírané do systému NUŠL z Národní lékařské knihovny a v průběhu navrženého experimentu byl srovnán výsledný věcný popis tvořený hesly PSH přiřazených na základě vytvořených mapování a věcný popis vytvořený automatickou indexací. Kromě toho byla řešena i možnost mapování autorských klíčových slov popisujících vysokoškolské kvalifikační práce v záznamech pocházejících z repozitářů spolupracujících vysokých škol.

Keywords: digital repositories, subject description, mapping, automatic indexing, National Repository of Grey Literature

Abstract:

The diploma thesis focuses on subject description unification methods in records aggregated from different sources in digital repositories, using the example of the National Repository of Grey Literature (NRGL). After presenting experiences with systems BASE and LASSO abroad, I describe the current situation in NRGL, where the automatic indexing is used to assign each record a unified subject heading from the Polythematic Structured Subject Heading System (PSH).

The thesis then presents how the MeSH thesaurus and Conspectus categorization scheme were mapped to PSH. These mappings were then applied to records from the National Medical Library. The aim of the experiment was to compare the subject description consisting of PSH subject headings created by automatic indexing, and the subject description created by mapping. In addition to that I explore the possibilities of mapping author keywords in records of academic theses.

Obsah

Seznam zkratek

API	rozhraní pro programování aplikací (Application Programming Interface)
BASE	Bielefeldský akademický vyhledavač BASE (Bielefeld Academic Search Engine)
DDT	Deweyho desetinné třídění
LASSO	LEAP Aggregated Search Service On-line
LEAP	London E-prints Access Project
MERLIN	Metadata Enrichment for Repositories in a London Institutional Network
MeSH	Medical Subject Headings
NK	Národní knihovna
NLK	Národní lékařská knihovna
NTK	Národní technická knihovna
NUŠL	Národní úložiště šedé literatury
OAI-PMH	Open Archives Initiative - Protocol for Metadata Harvesting
OCR	optické rozpoznávání znaků (optical character recognition)
PSH	Polytematický strukturovaný heslář

Předmluva

Téma této práce jsem se rozhodla zpracovat během práce v Národní technické knihovně, kde jsem od roku 2014 působila jako koordinátorka obsahu repozitáře NUŠL a zároveň jako správkyně Polytematického strukturovaného hesláře. Toto pracovní umístění mne přivedlo k zájmu o problematiku věcného popisu v digitálních repozitářích, jeho možnou automatizaci a optimalizaci a s tím související příležitosti a problémy. Výsledky této práce budou využity nejen v dalším vývoji systému NUŠL, ale již nyní jsou dílčí výstupy, jakými jsou např. vytvořená mapování, aplikovány do běžné praxe. Téma této práce bylo navrženo jako individuální téma závěrečné práce a prošlo příslušným schvalovacím procesem na Ústavu informačních studií a knihovnictví Filozofické fakulty Univerzity Karlovy v Praze.

Cílem práce bylo prozkoumat možnosti využití metod pro sjednocování věcného popisu v systému NUŠL a jejich srovnání. Předpokladem bylo, že mapování klasifikačních a kategorizačních schémat užívaných ve zdrojových záznamech bude pro sjednocování věcného popisu vhodnější a přesnější než již dříve implementovaná automatická indexace užívaná v NUŠL k tomuto účelu.

Úvodní část práce se zabývá již existující praxí sjednocování věcného popisu v zahraničních systémech agregujících záznamy z více zdrojů (kapitola 2) spolu se současným stavem řešení této problematiky v systému NUŠL (kapitoly 3-5). Podíl popisných částí musel být nakonec oproti původním předpokladům zvýšen, jelikož se ukázalo jako zásadní představit jednotlivé komponenty metod sjednocování věcného popisu, tedy systém NUŠL, heslář PSH, automatickou indexaci a mapovaná schémata. Následně byl na základě popsaných předchozích zkušeností vypracován plán experimentu, který tvoří druhou část práce.

V šesté kapitole je popsána tvorba mapování kategorizačního schématu Konspekt, tezauru MeSH a výsledky jejich aplikace na záznamy. Tato kapitola obsahuje také výsledky analýzy možnosti využití mapování u věcného popisu realizovaného pomocí volných klíčových slov. Závěrečná sedmá kapitola se zabývá srovnáním popsaných metod sjednocování na vzorku záznamů.

Pokud není uvedeno jinak, tabulky, grafy a schémata jsou dílem autorky této práce. Citace byly vytvořeny podle normy ISO 690. Přílohu tvoří CD s tabulkou obsahující vzorek záznamů, které byly použity ke srovnání jednotlivých metod sjednocování věcného popisu (viz kapitola 7).

Celkový počet znaků včetně mezer v hlavní části práce je 172 914, což po přepočtu představuje 96 normostran.

Závěrem bych ráda poděkovala PhDr. Evě Bratkové, Ph.D. za vedení práce, dále Janě Sloukové za pomoc s programováním a Marii Hamšíkové za jazykové korektury.

1. Úvod

Rozvoj výpočetní techniky v druhé polovině 20. století ovlivnil téměř veškeré obory lidské činnosti. Informační věda, je-li chápána v rozsahu definice jako „obor, který se zabývá člověkem zaznamenanou informací a zaměřuje se na komponenty komunikačního řetězce" (Bawden, 2012, s. 4), se samozřejmě musí zabývat některými změnami způsobenými tímto technickým rozmachem.

Z pohledu této práce jsou podstatné změny na začátku komunikačního řetězce (v českém prostředí se používá spíše termín informační cyklus, zde ho budeme chápat jako synonymní), tedy změny v oblasti vzniku dokumentů a následně nutně navazující změny ve zpracování dokumentů (těžiště práce) a jejich zpřístupňování.

Rozšíření počítačů a jejich následné propojení sítěmi (zvláště internetem) mělo za následek velký nárůst počtu digitálních dokumentů, a to digitalizovaných i tzv. „born digital", tedy dokumentů již vzniklých v digitální formě. Rozvoj digitálních dokumentů neznamenal pouze zvětšení objemu, ale šlo o změnu tak razantní, že lze hovořit o změně paradigmatu vnímání dokumentů a práce s informacemi a dokumenty jako takovými.

S rostoucím počtem digitálních dokumentů se instituce vypořádávaly i zakládáním digitálních repozitářů. Ty ve svých počátcích (1991 ArXiv.org) zajišťovaly přístup z jednoho místa k dokumentům, které by jinak byly obtížně dostupné a dohledatelné. Během dalších dvaceti let vzniklo ovšem takové množství repozitářů, že bylo obtížné až nemožné se v nich orientovat, a v důsledku toho začaly vznikat nejen registry těchto systémů, ale i služby zastřešující množiny těchto repozitářů a umožňující vyhledávání z jednoho rozhraní. Rozvoj těchto služeb ještě podpořil vývoj protokolu OAI-PMH, který umožňuje automatizované sklízení metadat z repozitářů.

Shromažďování metadatových záznamů z různých repozitářů i dalších zdrojů ovšem těmto službám přineslo potíže s věcným popisem, který se může mezi jednotlivými zdroji záznamů výrazně lišit, a to od záznamů prakticky bez věcných údajů, přes záznamy s volnými klíčovými slovy až po záznamy opatřené termíny z různých řízených slovníků. Obdobné problémy už dříve museli řešit katalogizátoři v souvislosti se sdílenou katalogizací, nicméně v oblasti repozitářů se objevují i další nové komplikace. Na rozdíl od přebírání klasických katalogizačních záznamů se v oblasti repozitářů a návazných služeb pracuje často s velkými

objemy nových záznamů najednou, není pravidlem dostupnost plného textu a tyto služby především nemívají tolik lidských zdrojů, aby mohl být odpovídajícím způsobem procházen a kontrolován či upravován každý jednotlivý záznam. Ke sjednocování záznamů na společnou úroveň věcného popisu nebo k užití stejného slovníku je tedy třeba využít nějaké automatizované či poloautomatizované metody.

2. Metody sjednocování věcného popisu v zahraničních systémech

V následující části budou na základě odborné literatury, dokumentace a vlastní práce s nimi popsány vybrané zahraniční systémy, které využívají nějaké formy sjednocování věcného popisu agregovaných záznamů. Vždy bude krátce představen systém jako takový, jeho zřizovatel a správce, zdroje a množství záznamů apod. Hlavní část se pak bude věnovat popisu metod věcného popisu a jeho sjednocení v systému.

Prvním představovaným systémem je systém BASE, ve kterém je dlouhodobě pracováno na sjednocování věcného popisu a zkušenosti z tohoto systému jsou dobře popsány a zveřejňovány. Druhým představovaným systémem je již neexistující systém LASSO, v rámci něhož probíhal projekt Merlin, který se stal inspirací i pro prvně zmiňovaný systém BASE.

2.1. Bielefeldský akademický vyhledavač BASE (Bielefeld Academic Search Engine)

BASE je víceoborový systém umožňující prohledávat indexovaný obsah (metadatové záznamy a v některých případech i plné texty) vědeckých informačních zdrojů, jako jsou digitální repozitáře nebo elektronické časopisy. Systém funguje již od června 2004 (Summann, 2004) a již od počátku jej spravuje Univerzitní knihovna Bielefeldské univerzity. V současné době obsahuje více než 93 milionů záznamů ze 4 351 zdrojů ze 104 zemí světa[1]. Systém původně využíval software FAST, ale v roce 2011 došlo k jeho migraci na současné řešení používající vyhledávací rozhraní VuFind a platformu Apache Solr (Universitätsbibliothek Bielefeld, 2004c).

2.1.1. Zdroje obsahu

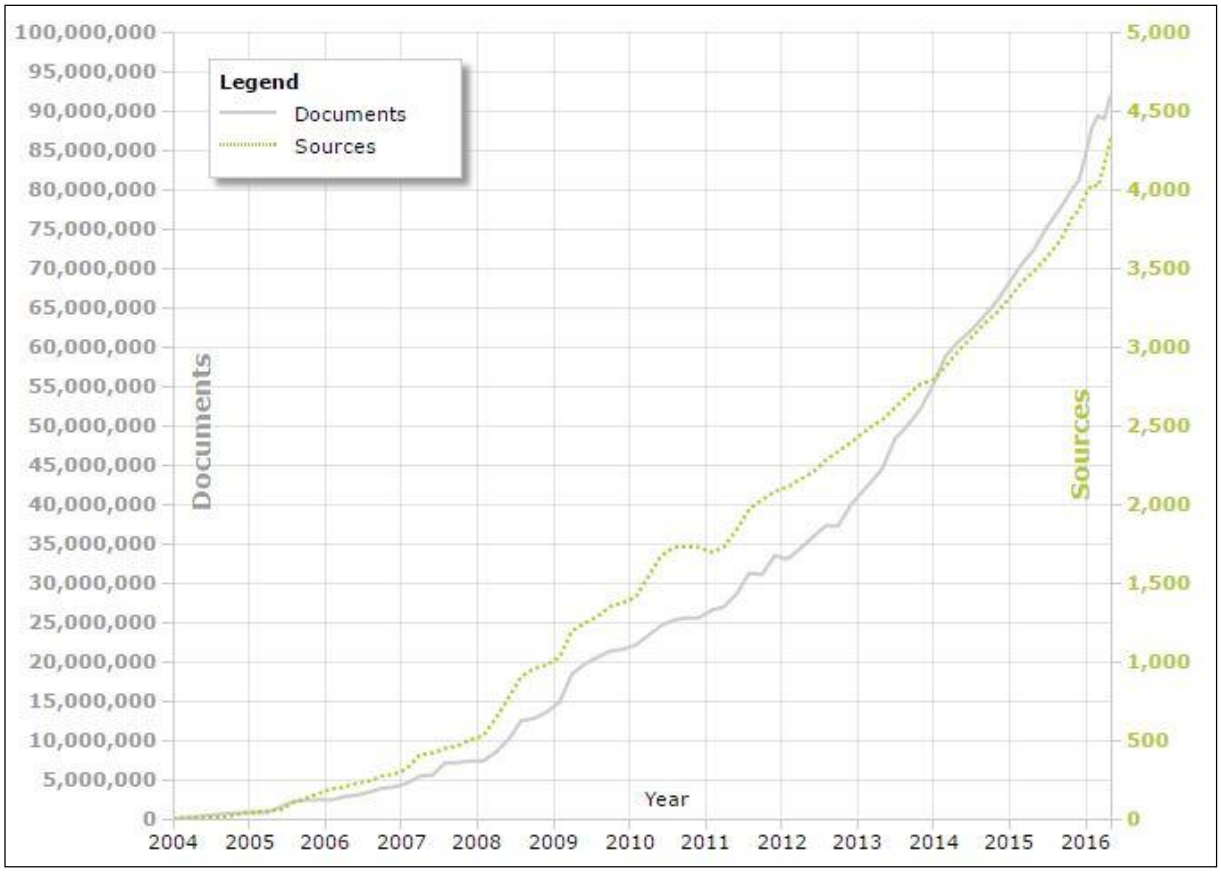
BASE uvádí, že zpracovává metadatové záznamy dokumentů ze „všech druhů akademicky relevantních repozitářových serverů, které používají OAI-PMH k poskytování svých metadat“ (Universitätsbibliothek Bielefeld, 2004b). Systém neumožňuje jiný způsob předávání dat než skrze OAI-PMH protokol. BASE provozuje OAI-PMH validátor OVAL, který umožňuje zkontrolovat, zda OAI-PMH zdroje splňují všechny požadavky systému.

[1] Údaje platné k 30. květnu 2016.

Jednotlivé zdroje jsou vytipovávány sledováním registrů repozitářů (OpenArchives[2], ROAR[3], OpenDOAR[4]) a registrů specializovaného softwaru pro repozitáře (DSpace[5], OJS - Open Journal Systems) (Universitätsbibliothek Bielefeld, 2004a). Zároveň je na stránkách BASE webový formulář umožňující správcům repozitářů či elektronických časopisů navrhnout zařazení jejich zdroje do systému. Výběr zdrojů není nijak oborově, územně ani jazykově omezen.

2.1.2. Sjednocování věcného popisu v systému BASE

Široký záběr BASE neomezující se tematicky, jazykově ani geograficky měl za následek rychlý růst metadatové databáze již od samého počátku fungování systému. Po prvním roce činnosti obsahoval více než 2 300 000 záznamů, za další rok již 7 milionů a nyní po dvanácti letech již 93 milionů záznamů (růst viz graf na obr. 1).



Obrázek 1: Růst počtu zdrojů a dokumentů v systému BASE (About BASE: Statistics, 2016)

[2] https://www.openarchives.org/Register/BrowseSites
[3] http://roar.eprints.org/
[4] http://opendoar.org/
[5] https://wiki.duraspace.org/display/DSPACE/OaiInstallations

S takto rychlým nárůstem v řádech milionů záznamů rozhodně nelze případné sjednocení věcného popisu provádět jinak než automatizovaně. Další překážkou jednotného věcného popisu je víceoborovost systému vzhledem k odlišným zvyklostem ohledně užívaných klasifikačních či jiných schémat v jednotlivých oborech. Nejednotnost je pak podpořena i velkým počtem zemí, z nichž jednotlivé zdroje pochází, protože mohou využívat vlastní národní pořádací systémy.

Kromě těchto překážek vycházejících ze základních principů systému BASE byly v Bielefeldu definovány další problémy s věcným popisem v Dublin Core, který je základním formátem OAI-PMH protokolu:

1. Při převodu metadat z nějakého detailnějšího formátu na DC dochází ke ztrátě informací.
2. Pole věcného popisu se mohou opakovat, aniž by měla nějaký další kvalifikátor (přestože obsahují např. údaje z jiného klasifikačního schématu než v předchozím poli stejného jména apod.).
3. Neexistují pevná pravidla udávající, co a jak se má v prvku dc:subject uvádět. Jako obsah tohoto prvku se tak objevují deskriptory z různých tezaurů, volná klíčová slova, slova z jiných prvků (autor, název), notace z různých klasifikačních schémat i další (Lösch, 2009).

Všechny tyto překážky znesnadňují a znepřesňují vyhledávání v systému. Lösch navrhuje dva směry jak tyto problémy řešit, a to standardizaci, která je ovšem zvláště napříč obory, jazyky i zeměmi prakticky neuskutečnitelná (nebo alespoň ne v krátkém časovém horizontu), a „automatickou homogenizaci věcného popisu", pro niž se v BASE následně rozhodli.

2.1.3. Projekt Automatického obohacování OAI metadat

Dvouletý projekt s názvem „Automatické obohacování OAI metadat s pomocí metod počítačové lingvistiky a vývoj služeb pro obsahové propojování repozitářů[6]" probíhal od října 2009 do konce září 2011. Projekt byl financovaný německou výzkumnou společností DFG (Deutsche Forschungsgemeinschaft) a spolu s Bielefeldskou univerzitní knihovnou na něm spolupracovala Laboratoř textových technologií Univerzity Johanna Wolfganga Goetheho

[6] V originále „Automatische Anreicherung von OAI-Metadaten mit Hilfe computerlinguistischer Verfahren und Entwicklung von Services für die inhaltsorientierte Vernetzung von Repositorien".

ve Frankfurtu nad Mohanem a Oddělení automatického zpracování řeči Lipské univerzity (Universitätsbibliothek Bielefeld, 2010).

2.1.3.1. Cíle projektu

Správci systému BASE se snaží směřovat vývoj směrem k sémantickému vyhledávání, které nebude závislé na slovní reprezentaci, ale bude založené na vyhledávání podle pojmů. V rámci tohoto projektu se ke svému cíli měli v úmyslu přiblížit obohacením předmětového popisu záznamů o notace Deweyho desetinného třídění (DDT). Tato data mělo být možné využít i v dalších repozitářích a vyhledávačích k přesnějšímu vyhledávání. Jedním z plánovaných a později skutečně realizovaných využití takto obohacených záznamů bylo vytvoření prohlížení (browsing) záznamů podle oborů v systému BASE.

2.1.3.2. Průběh projektu

Projekt přiřazování notací DDT k záznamům lze rozdělit do několika hlavních částí a procesů:

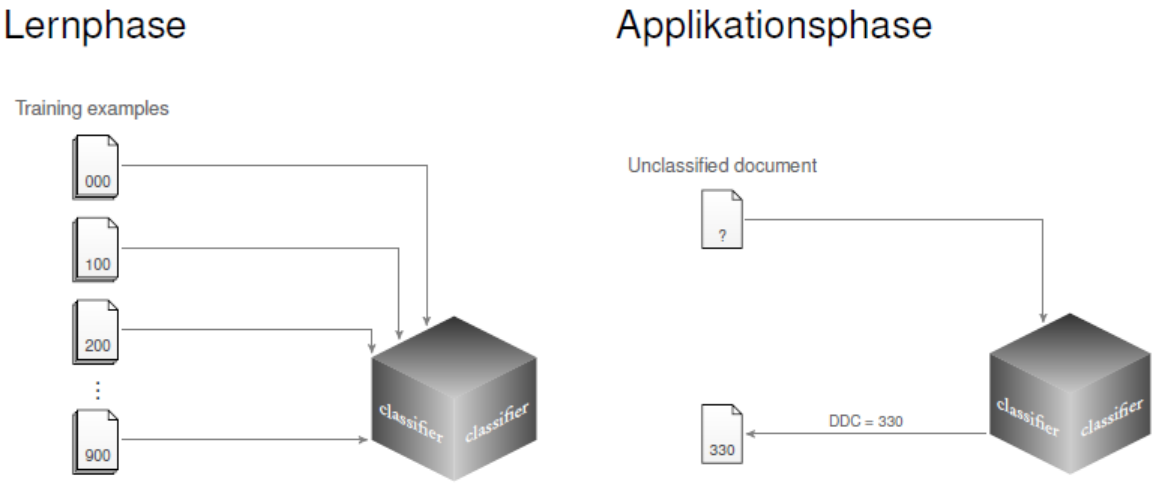
1) Učící fáze

- Výběr množiny záznamů k vytvoření korpusu
- Sjednocení věcného popisu u vybraných záznamů
- Vytvoření modelu strojového učení

2) Aplikační fáze

- Přiřazení relevantních notací DDT k dalším záznamům pomocí automatické klasifikace na základě trénovacích dat

Na obrázku 2 je zjednodušené schéma naznačující vytváření modelu na základě trénovacích dat v učící fázi a klasifikaci dokumentů v aplikační fázi.



Obrázek 2: Fáze projektu v BASE (Lösch, 2011)

Z hlediska této práce je zajímavé, že během projektu v BASE byly využity oba přístupy, které budou později porovnávány ve druhé části práce na případu Národního úložiště šedé literatury, tedy mapování klasifikačních schémat i automatická indexace. K mapování bylo přistoupeno v první učící fázi a jeho výsledky posloužily k vytvoření modelu strojového učení využitého v druhé, aplikační fázi projektu.

V učící fázi bylo nutné vytvořit korpus tréninkových dat pro indexér. Indexér měl mít po skončení této fáze k dispozici korpus dat obsahující text dokumentu a jeho OAI záznam včetně alespoň jedné notace DDT. Indexér pak takto „naučené" informace v druhé fázi využil k přiřazení notací DDT k neklasifikovaným dokumentům (viz obr. 2).

1) Učící fáze

Výběr množiny záznamů k vytvoření korpusu

Jako zdroj záznamů vhodných pro vytvoření korpusu posloužil systém BASE, který měl v roce 2009, kdy projekt začínal, již na 21 milionů záznamů z více než 1 300 zdrojů (About BASE: Statistics, 2015). Záznam musel splňovat tři podmínky, aby mohl být zařazen do korpusu:

1. Do korpusu mohly být zařazeny pouze záznamy s dokumenty v anglickém a německém jazyce.
2. Záznam musel odkazovat na elektronicky dostupný plný text dokumentu. Ten musel být k dispozici volně ke stažení ve strojově čitelném formátu (například i PDF dokument s OCR).

3. Musel existovat způsob, jak přiřadit notaci DDT. Zařazeny tedy byly jak záznamy, které již přímo obsahovaly notaci DDT, tak záznamy s notacemi nebo termíny z jiných klasifikačních schémat, u nichž bylo možné jejich věcný popis převést na DDT.

Práce se záznamy

V rámci projektu se pracovalo s metadaty agregovanými pomocí protokolu OAI-PMH. Vybráno bylo metadatové schéma OAI DC, které je uzpůsobeno pro předávání protokolem OAI-PMH a odpovídá nekvalifikovanému schématu Dublin Core. Jedná se o jediné metadatové schéma, které musí být dle specifikace OAI-PMH 2.0 povinně přístupné v každém repozitáři provozujícím OAI-PMH; tato specifikace zároveň pevně stanovuje i prefix „oai_dc" označující toto schéma v protokolu. OAI DC je pak vhodnou volbou, jelikož se bude vyskytovat ve všech harvestovaných repozitářích a navíc pod stejným označením, což jsou výhody, které jiná (byť podrobnější) metadatová schémata nemají. Záznamy jsou ve formátu XML.

V první fázi bylo nutné určit použité klasifikační schéma v záznamu, jelikož s údaji z různých klasifikačních schémat bylo třeba nakládat odlišně. V rámci OAI DC nejsou používány kvalifikátory, které by umožnily snadné určení použitého klasifikačního schématu, a proto musel být vyvinut způsob jejich rozlišování. V některých případech sice lze spoléhat na politiku věcného popisu daného zdrojového repozitáře, problém ale nastal v případech, kdy repozitář využívá více schémat (ať už je sám přiřazuje či agreguje další zdroje), nebo dokonce u záznamů obsahujících věcné údaje z více klasifikačních schémat. Bylo tedy nutné vyvinout automatizovaný způsob, který umožnil určit schéma na základě struktury obsahu prvku dc:subject, v němž se v OAI DC nacházejí informace věcného popisu.

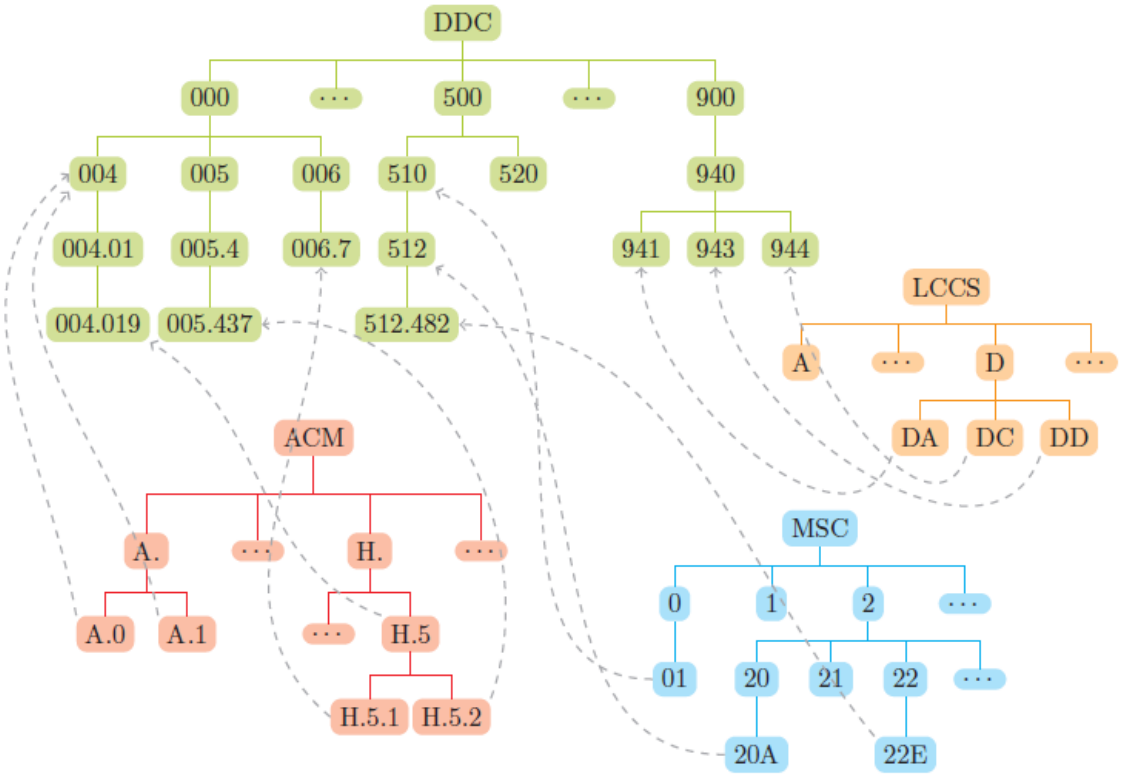
Po určení schématu byla ze záznamů, které již měly přidělenou DDT notaci ze svých původních repozitářů, tato hodnota jednoduše převzata. U záznamů, v nichž byla využita jiná klasifikační schémata, musel být zajištěn jejich převod na DDT.

Pro převod na DDT byla vytvořena řada mapování s DDT jako cílovým schématem. Idea tohoto mapování je znázorněna na obr. 3. Byla vytvořena částečná mapování pro tato schémata:

- Klasifikační systém výpočetní techniky ACM (ACM Computing Classification System)
- Kategorie ArXiv.org
- Nizozemská základní klasifikace (Nederlandse Basisclassificatie)
- Joint Academic Coding System

- Klasifikace časopisu Journal of Economic Literature (JEL Classification)
- Klasifikační schéma JITA (JITA Classification Schema)
- Klasifikace Kongresové knihovny (Library of Congress Classification)
- Matematická předmětová klasifikace Americké matematické společnosti (Mathematics Subject Classification)
- Klasifikační schéma fyziky a astronomie PACS (The Physics and Astronomy Classification Scheme)
- Regensburská klasifikace (Regensburger Verbundklassifikation)

Nebyla vytvářena kompletní mapování mezi DDT a jednotlivými schématy, ale spíše se přihlíželo k tomu, z jakých oborů pochází záznamy, které je třeba převést na DDT. Mapování v tomto případě bylo skutečně jen způsobem získání určitého množství dat, které pak umožní plnou automatizaci věcného popisu, proto nebylo třeba provádět ho kompletně.



Obrázek 3: Idea mapování klasifikačních schémat na DDT (Waltinger, 2010)

Hlavním úkolem této části je přiřadit každému záznamu vybranému k vytvoření korpusu alespoň jednu notaci DDT. Přiřazená notace DDT je pak spolu s dalšími údaji získanými zpracováním plného textu (haš[7] plného textu) přidána do záznamu v XML. Takto obohacené

[7] Zapisuje se často i jako hash.

záznamy byly následně ukládány do SQL databáze dostupné přes rozhraní HTTP, jež bylo zvoleno pro snadnější přístup podle věcné klasifikace (Lösch et al., 2011, s. 4).

Zpracování plných textů

Druhou hlavní složkou korpusu BASE je plný text vybraných záznamů. Pro každý repozitář je nastaveno, kde v záznamu hledat odkaz na plné texty, aby je bylo možné stáhnout. Stažený dokument je pak převeden do prostého (holého) textu bez formátovacích znaků a jako prostý textový dokument je i uložen a následně je ověřen jazyk daného dokumentu. V BASE považovali tento krok za nezbytný, jelikož ne všechny zdrojové repozitáře uváděly prvek dc:language s údajem o jazyce dokumentu a zároveň je tu vždy prostor pro chybné označení v záznamu. Tímto se tedy zamezí tomu, aby chybně jazykově zařazený dokument ovlivnil přesnost automatické klasifikace dalších tisíců záznamů (Lösch et al., 2011, s. 3).

Celý text dokumentu je pak pomocí hašovacího algoritmu MD5 převeden na 32 znaků šestnáctkové soustavy a tento řetězec (nazývaný haš) pak slouží jako unikátní identifikátor textu a je spolu s notací DDT přiřazen do XML záznamu. Haš v záznamu zajišťuje jeho propojení s plným textem příslušícím k záznamu. Běžně se zmíněný algoritmus MD5 používá pro kontrolu integrity dokumentu, ale v případě tohoto projektu byl použit především jako ochrana před zařazením duplicitních dokumentů do korpusu. Při přidávání nového textu byly vždy porovnávány haše, protože dva identické záznamy by zbytečně narušily přesnost následné automatické indexace (Lösch et al., 2011, s. 4).

Text dokumentů bylo pro potřeby vytvoření učícího modelu automatické indexace nutné dále zpracovat. Texty prošly segmentací, při níž byly rozděleny na jednotlivá slova a slovní spojení, lemmatizací zajišťující převod slov na základní gramatický tvar, gramatickým značkováním, což je proces „přiřazení (symbolů) značek slovních druhů každému výskytu slova v korpusu" (Pala, 1996), rozpoznáváním pojmenovaných entit a eliminací stop slov. (Lösch, 2009).

Korpus

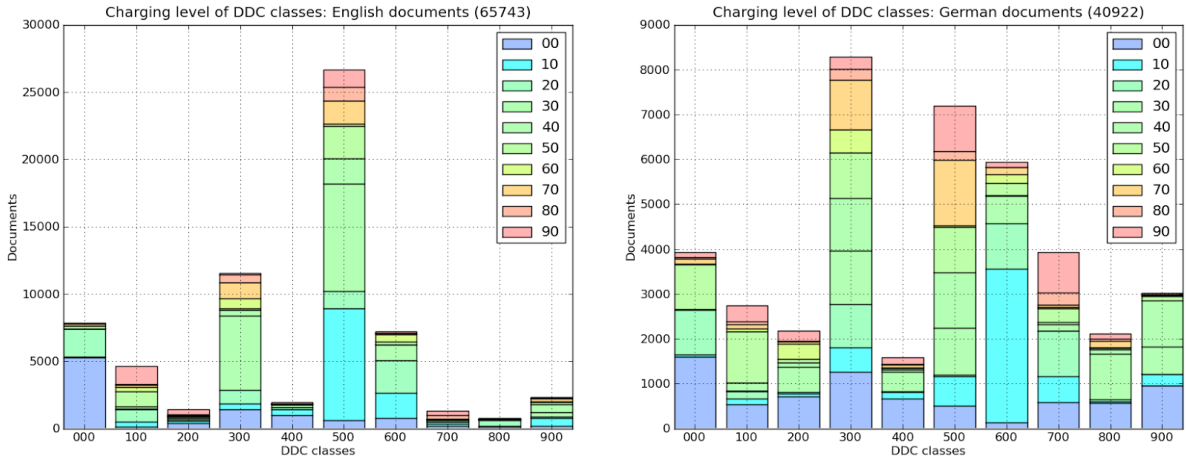
K vytvoření korpusu bylo využito 52 905 plných textů v anglickém jazyce a 37 228 textů v německém jazyce. Tyto dokumenty pocházely ze 101 repozitářů, jejichž záznamy jsou harvestovány do systému BASE. Pokud jde o objem dat, pokud bereme v úvahu nekomprimované texty v kódování UTF-8, byla celková velikost použitých OAI záznamů 439 MB a velikost zpracovaných plných textů 13 GB (Lösch et al., 2011, s. 4).

Každý záznam musel být zařazen alespoň do jedné kategorie první úrovně a celkově bylo možné zařazovat záznamy do prvních tří úrovní hierarchie DDT. Jelikož bylo cílem využít korpus pro automatickou indexaci, musely být stanoveny limity pro množství dokumentů v jednotlivých třídách, aby nedošlo k narušení rovnováhy v dalších procesech, kde by třídy třeba až s několikanásobně větším počtem dokumentů narušily přesnost automatické indexace. Maximální limity byly tedy stanoveny na 10 000 dokumentů pro nejvyšší úroveň DDT, 1 000 dokumentů pro druhou úroveň a 100 dokumentů pro třetí úroveň DDT (Lösch et al., 2011, s. 6). Po naplnění kapacity dané třídy už do ní nebyly další dokumenty zařazovány.

Minimální limity pro třídy jednotlivých úrovní stanoveny nebyly a na druhé a třetí úrovni hloubky DDT zůstaly nepokryty. Tuto nerovnováhu autoři vysvětlují nekonzistencemi ve struktuře DDT a také problémem nedostatečného množství volně dostupných dokumentů z některých vědních oborů. Zatímco se získáváním dokumentů v oborech, kde je silná tradice publikování v rámci otevřeného přístupu (například fyzika), nebyly problémy a naopak musel být právě kvůli těmto oborům stanoven zmíněný horní limit pro počet dokumentů ve třídách, u oborů tolik nenakloněných otevřenému přístupu nastal se získáním dostatečného množství dokumentů problém a v korpusu jich bylo v zpočátku méně, jak je vidět v tabulce na obr. 4 znázorňující množství dokumentů podle zařazení do první úrovně. Graf na obr. 5 ukazuje množství dokumentů v angličtině a němčině zařazených do druhé úrovně DDT. Tento stav měl být později napraven a třídy doplněny o dokumenty - ke konci projektu autoři uváděli, že korpus obsahuje zhruba 100 000 dokumentů (Lösch, 2011).

DDC	English	German
000 Computer Science, information & general works	6847	3778
100 Philosophy & psychology	3536	2169
200 Religion	1123	1973
300 Social sciences	10948	8075
400 Language	1682	1297
500 Science	23989	6969
600 Technology	6669	5874
700 Arts & recreation	1280	3823
800 Literature	740	2063
900 History & geography	2226	2863

Obrázek 4: Počet dokumentů zařazených do jednotlivých DDT kategorií (Lösch et al., 2011, s. 6)



Obrázek 5: Rozdíly klasifikací anglických a německých dokumentů (Lösch, 2011, s. 14)

Vytvoření modelu strojového učení

Na základě výsledků zpracování textů a záznamů proběhl výběr jejich rysů neboli atributů, které byly použity pro konstrukci vektorového modelu automatické indexace. Tvůrci projektu zvolili jako metodu strojového učení Support Vector Machines, která se do češtiny také někdy překládá jako metoda podpůrných vektorů.

2) Aplikační fáze

V aplikační fázi byla s využitím korpusu a vektorového modelu vytvořeného v učící fázi přidělena notace DDT do dalších záznamů s pomocí automatické indexace.

Bez plných textů

V rámci projektu byla testována i varianta automatické indexace nikoli s využitím plných textů, ale pouze bibliografických záznamů.

K indexaci byly vybrány záznamy v anglickém či německém jazyce, které popisovaly vědecké dokumenty, články a prezentace v rozsahu do 100 stran. Původně měly být zařazeny pouze záznamy, jejichž pole „popis“ (dc:description) bylo větší než 100 bytů, nicméně tato hodnota musela být snížena na 30 bytů, aby bylo zajištěno dostatečné množství záznamů k experimentování. Tato kritéria nakonec splňovalo 20 813 záznamů v angličtině a 37 769 záznamů v němčině. Ke klasifikaci byly použity pouze prvky z OAI záznamů obsahující název, předmět a popis (title, subject, description) (Waltinger et al., 2009, s. 33).

Jako první krok bylo u každého záznamu automaticky zjištěno, zda již neobsahuje notaci DDT. Pak byly postoupeny k předzpracování, kdy byl rozlišen jazyk záznamu a proběhla lemmatizace obsahu vybraných polí OAI záznamů. Nakonec byly vytvořeny dva soubory dat

(datasety) předzpracovaných německých a anglických záznamů označených notacemi DDT (Waltinger et al., 2009, s. 33).

K automatické indexaci byla pak rovněž použita metoda podpůrných vektorů, přičemž pro každou třídu byl vytvořen vlastní model. Přidělovány byly opět notace DDT do třetí úrovně, ale v tomto případě mohlo dojít i k nepřiřazení záznamu do žádné kategorie.

Tento experiment byl vyhodnocen tak, že přesnost automatické indexace na nejvyšší úroveň DDT je uspokojivá - výsledek statistické analýzy vychází pro anglické záznamy na hodnoty 0,81 F1 míry[8] a 0,79 F1 míry pro němčinu. Na druhé úrovni hloubky DDT se pak jedná o horší hodnoty 0,74 (němčina) a 0,63 (angličtina) a na třetí úrovni už jen 0,61 (němčina) a 0,62 (angličtina), ovšem ve výsledcích se odráží i to, že pro druhou a třetí úroveň nebyl dostatek trénovacích dat (Waltinger et al., 2009, s. 35-37).

2.1.3.3. Výsledky projektu

Před zahájením projektu mělo v BASE přiřazeno notaci DDT pouze 429 496 záznamů, což při tehdejší velikosti systému tvořilo zhruba 1,4 % databáze. Po skončení projektu mělo tuto notaci již 1 753 712 záznamů a toto číslo dále rostlo (Universitätsbibliothek Bielefeld, 2011). Dnes už se indexace nezaměřuje pouze na záznamy s volně přístupnými dokumenty a notace mají i záznamy bez plných textů, takže se dá předpokládat, že byla zprovozněna i klasifikace nezávislá na plných textech.

Nově mohou tato data uživatelé využít v rozhraní prohlížení BASE (Browsing), kde mohou procházet dokumenty podle zařazení do kategorie DDT (rozhraní prohlížení je vyobrazeno na obr. 6).

Obrázek 6: Rozhraní prohlížení v BASE

[8] Míra F1 je harmonizovaná střední hodnota kombinace přesnosti a úplnosti.

Výstupem projektu je zároveň i automatický indexér, který je dostupný buď prostřednictvím API[9], nebo skrze webové rozhraní, v němž je možné otestovat klasifikaci části textu, celého souboru PDF či webové stránky. Tento indexér pracuje pouze s texty v angličtině a němčině (sám umí automaticky rozlišit tyto dva jazyky).

K dispozici jsou i mapování mezi DDT a dalšími klasifikacemi, která byla vytvořena při budování korpusu, nicméně většinou se jedná o mapování hodně hrubá a nekompletní (v některých případech je namapována jen část, která byla potřebná pro projekt).

[9] API je zkratka pro „Application Programming Interface“, rozhraní pro programování aplikací.

2.2. LASSO a projekt MERLIN

Systém LASSO (zkratka vychází z anglického názvu LEAP Aggregated Search Service Online) byla agregační služba harvestující záznamy z repozitářů členů konsorcia SHERPA-LEAP. V rámci této služby byl uskutečněn projekt MERLIN, který měl za úkol prozkoumat a implementovat cenově efektivní způsob automatického věcného popisu v repozitářích.

2.2.1. SHERPA-LEAP

Konsorcium londýnských vysokoškolských institucí SHERPA-LEAP (London E-prints Access Project) bylo založeno v roce 2004 (Moyle et al., 2007, s. 1) s cílem podporovat vznik, vývoj a naplňování repozitářů elektronických tisků v rámci federální Londýnské univerzity. Z původních sedmi členů postupně narostl počet až na 13 institucí lišících se velikostí i zaměřením (nejednalo se o všechny instituce spadající pod Londýnskou univerzitu). Ve vedení konsorcia stála University College London (UCL), která rovněž spravovala centrální server (Moyle et al., 2007, s. 3).

Jako software pro repozitáře si zakládající členové vybrali open source EPrints (zejména kvůli dostupnosti technické podpory od organizace SHERPA a předpokládané rychlosti a jednoduchosti implementace). V pozdějších fázích projektu pak dvě členské instituce provozovaly své lokální instalace DSpace, jeden člen využíval vlastní instalaci zmíněného systému EPrints a deset členů konsorcia používalo sdílenou instalaci EPrints na serveru UCL. V rámci sdílené instalace byly jednotlivé repozitáře nastaveny tak, aby si funkcionalitu i rozhraní každá instituce mohla konfigurovat sama (Moyle et al., 2007, s. 3).

V rámci repozitářů konsorcia nebyla sjednocena typologie dokumentů a repozitáře nepoužívaly žádné společné klasifikační schéma věcného popisu (tato varianta byla zavržena už při vzniku konsorcia LEAP vzhledem k rozpětí oborů pokrývaných institucemi konsorcia) (Moyle, 2011, s. 6). Instituce si mohly samostatně rozhodnout i o metadatových schématech, podporovaných formátech souborů, akviziční politice i o nastavení procesu ukládání dokumentů (a to i v případě sdílené kopie systému EPrints) (Moyle et al., 2007, s. 3).

Po vytvoření repozitářů a přidání dalších členů konsorcia byla dalším krokem implementace vyhledávacího rozhraní nad repozitáři konsorcia LASSO.

2.2.2. LASSO

Služba LASSO byla v roce 2008 vyvinuta na UCL, aby poskytovala rozhraní pro vyhledávání nad repozitáři členů konsorcia SHERPA-LEAP. Služba LASSO harvestovala jednou denně pomocí protokolu OAI-PMH metadata ve formátu OAI Dublin Core z 9 repozitářů konsorcia. V roce 2008 měla umožňovat vyhledávání v 13 000 záznamech (MERLIN Project Proposal, 2008, s. 2) a v roce 2011 již v nejméně 15 000 záznamech.

Po naharvestování probíhala normalizace sklizených dat, a to kvůli odlišnostem s užitím Dublin Core v jednotlivých zdrojových repozitářích (pomocí mapování například došlo ke sjednocení typologie dokumentů na jednotnou sadu) (MERLIN Project Proposal, 2008, s. 2).

Vyhledávat bylo možné podle autora, názvu, slov abstraktu a klíčových slov dodaných původním repozitářem (pokud jsou údaje k dispozici). Zároveň bylo možné výsledky seřadit podle názvu, instituce nebo roku vydání (MERLIN Project Proposal, 2008, s. 2).

2.2.3. MERLIN: Metadata Enrichment for Repositories in a London Institutional Network

Projekt Obohacení metadat z repozitářů v londýnské institucionální síti SHERPA-LEAP, pro který se používá zkratka MERLIN odvozená od jeho anglického názvu, probíhal od května 2009 do února 2011. Hlavním cílem tohoto projektu bylo pomocí již existujících dostupných technik analýzy textu (text mining) obohacovat záznamy v databázi LASSO. Projekt byl financován britskou organizací JISC v rámci grantové výzvy 12/08 A1 Automatická generace metadat a text mining (MERLIN Project Proposal, 2008, s. 1).

2.2.3.1. Cíle projektu

Projekt měl při podávání tyto cíle:

- S použitím text miningového nástroje TerMine obohatit záznamy v databázi LASSO o klíčová slova automaticky vygenerovaná ze zdrojových repozitářů.
- Vytvořit design a implementovat změny uživatelského rozhraní LASSO tak, aby odrážel relevantní klíčová slova na úrovni kolekcí i podkolekcí.
- Zapracovat relevantní klíčová slova do rozhraní pokročilého vyhledávání.
- Provést uživatelské testování rozhraní s obohacenými záznamy, aby bylo uzpůsobeno novým možnostem užití.
- Do evaluace nového rozhraní měli být zahrnuti i koncoví uživatelé.

- Využít poznatky z projektu HILT (High Level Thesaurus) k vytvoření stromu klasifikace předmětových hesel získaných pomocí text miningu.
- Vytvořit techniku obohacování MERLIN záznamů tak, aby nebyla závislá na konkrétní technické platformě a stala se z ní znovupoužitelná a open source webová aplikace. (MERLIN Project Proposal, 2008, s. 1)

2.2.3.2. Průběh projektu

Samotný projekt byl rozdělen do několika „pracovních balíčků", z nichž jsou pro tuto práci důležité hlavně druhý a čtvrtý:

1. Projektový management: Zabezpečuje plánování projektu a dohled nad jeho celkovým průběhem. Zajišťuje efektivní práci ostatních pracovních balíčků a zaručuje, že všechny výstupy budou hotovy včas a v rámci rozpočtu.
2. Extrakce termů: Zajišťuje vytvoření a testování vlastní text miningové metodologie, vlastní extrahování pojmů z plných textů pomocí specializovaného softwaru, vážení těchto termínů.
3. Integrace: Integrace termínů vyextrahovaných v balíčku 2 do rozhraní LASSO, úprava uživatelského rozhraní.
4. Vytvoření strukturované navigace: Za použití nástrojů vyvinutých v projektu HILT je vytvořen strom klasifikace založený na termínech vyextrahovaných ve 2. pracovním balíčku.
5. Evaluace: Evaluace nového rozhraní a přínosů obohacených záznamů oproti původní základní verzi.
6. Znovupoužitelnost: Příprava nástrojů, které bude možno použít v jiných systémech, budou nezávislé na konkrétní platformě a k dispozici jako open source. (MERLIN Project Proposal, 2008, s. 4-5)

Extrakce termů

Tento pracovní balíček se skládal z několika základních částí, které budou následně rozebrány:

a. Shromáždění plných textů ze zdrojových repozitářů
b. Převod textů dokumentů na prostý text bez formátování
c. Zpracování textu
d. Uložení vybraných termínů do databáze

Shromáždění plných textů ze zdrojových repozitářů

V rámci každodenního harvestování záznamů do systému LASSO začaly být záznamy testovány. Bylo kontrolováno, zda obsahují prvek dc:format, jehož přítomnost měla označovat dostupnost plného textu dokumentu v repozitáři. Pokud je při této kontrole nalezen plný text, je záznam označen a v další fázi je pak u označených záznamů automaticky stažen ze zdrojového repozitáře příslušný dokument. Celkově bylo takto označeno 15 000 záznamů (Moyle, 2011, s. 11).

V tomto kroku museli zpracovatelé projektu řešit problémy s různým užitím prvků Dublin Core, kdy se lišily prvky, do nichž byla v jednotlivých repozitářích ukládána URL plného textu (dc:identifier, dc:relation), a některé repozitáře do OAI DC tuto informaci nepředávaly vůbec. V této fázi se proto přešlo ke stahování metadat ve formátu METS, v němž byla správná URL lépe vyznačena.

Zároveň byl problém i s případným embargem na zpřístupnění plných textů, protože takové adresy navedly nástroj pro stažení nikoli k plnému textu, ale k obrazovce vyžadující přihlašovací údaje pro vstup do systému (Moyle, 2011, s. 8).

Převod textů dokumentů na prostý text

K této akci byly využity hned tři programy, kdy každý byl určen pro převod z jiných formátů. K převodu formátů kancelářské sady MS Office byl využit software Antiword, PDF dokumenty pomocí programu Xpdf (konkrétně pdftotext) a soubory ostatních formátů byly převáděny programem Java OpenDocument Converter.

Klasickým problémem byly PDF soubory bez OCR, které pak neprošly úspěšně další fází. U formátu PDF ovšem ještě před další fází docházelo k předzpracování, kdy byla identifikována a odstraněna data, která by při dalším zpracování působila nepřesnosti: tabulky, seznamy použité a doporučené literatury a různé šablony dokumentů (Moyle, 2011, s. 9).

Zpracování textu

K úvodnímu zpracování textu se v projektu používal rozdělovač vět, který byl (stejně jako TerMine) vyvinut Národním centrem pro text mining NaCTeM. Tento rozdělovač pomocí nastavených pravidel rozdělil texty na věty a odstavce a takto připravené byly dále posunuty do aplikace TerMine. TerMine rozčlenil texty na termíny a „aplikoval na ně statistickou analýzu k odvození vah ukazujících relativní důležitost daného termínu v dokumentu" (Moyle, 2011, s.

9). Z celkového množství 15 000 záznamů označených jako mající plný text jich zpracováním prošlo 10 000 a bylo z nich vyextrahováno 650 000 unikátních termínů (Moyle, 2011, s. 11).

Pokud to bylo nutné, byly následně provedeny ještě další úpravy termínů, jako například krácení příliš dlouhých termínů, v nichž se opakovala slova. Proběhl také pokus o druhotné vážení, kdy by termíny z názvu dokumentu a abstraktu měly vyšší váhu, ale nedošlo k převodu do praxe (Moyle, 2011, s. 11).

Uložení vybraných termínů do databáze

Výsledné termíny byly uloženy do SQL databáze, která byla rozdělena do dvou tabulek, z nichž jedna obsahovala termíny a druhá údaj o propojení mezi termíny a plnými texty (těchto propojení bylo přibližně 1 milion) a o váhách jednotlivých termínů (Moyle, 2011, s. 10-11).

Vytvoření strukturované navigace

Systém byl nakonfigurován tak, aby komunikoval pomocí protokolu SOAP[10] se serverem HILT. Po vyhledání termínu v databázi LASSO byla vždy prohledána místní databáze termínů a zároveň vyslán dotaz na server HILT, kde byl termín dohledán ve vybraném tezauru (LASSO využívalo tezaurus UNESCO); zpátky byly zaslány údaje o nadřazených, podřazených a souvisejících termínech. Pokud se následně uživatel přepnul do „módu thesaurus“, zpřístupnily se mu i tyto termíny (Moyle, 2011, s. 11). Veškerá data zaslaná ze severu HILT byla ukládána v LASSO.

2.2.3.3. Výsledky

Po skončení projektu měla mít databáze 10 000 záznamů obohacených o klíčová slova extrahovaná z plných textů dokumentů, další možnosti nakládání s termíny skrze službu HILT a nové rozhraní umožňující práci s těmito termíny. Služba LASSO však již dnes není v provozu a v provozu nejsou ani stránky konsorcia SHERPA-LEAP, nicméně vyhledávácí rozhraní MER, které bylo v rámci projektu MERLIN vytvořeno, je stále přístupné na Google code[11].

[10] Protokol zajišťuje okamžitou výměnu zpráv mezi webovými službami ve formátu XML.
[11] https://code.google.com/p/jisc-merlin/

3. Národní úložiště šedé literatury

Národní úložiště šedé literatury (též známé pod zkratkou NUŠL) je systém spravovaný Národní technickou knihovnou, jehož cílem je shromažďovat a zpřístupňovat šedou literaturu z oblasti vědy, výzkumu, vzdělávání a kultury vznikající na území České republiky.

Tato kapitola je zařazena pro podrobnější seznámení se systémem NUŠL, které je potřebné pro pochopení již provedených pokusů o sjednocení věcného popisu (kapitola 5) i pro lepší uchopení experimentu prováděného v rámci této práce. Vzhledem k zaměření experimentu je pozornost při představování systému zaměřena na zdroje záznamů a jejich podobu. Popsána je typologie dokumentů, zdrojové instituce záznamů a jednotlivá pole - povinná a nepovinná pro různé kategorie záznamů. Tento popis má za úkol ilustrovat rozmanitost záznamů. Kvalita záznamů kolísá, jelikož jsou vytvářeny mnoha lidmi v rozličných systémech a za různými účely. Zároveň jsou to záznamy různých druhů dokumentů pocházející z mnoha oborů.

NUŠL byl vytvořen v rámci projektu „Digitální knihovna pro šedou literaturu - funkční model a pilotní realizace" financovaném Ministerstvem kultury, na němž v letech 2008 až 2011 spolupracovaly Národní technická knihovna[12] a Vysoká škola ekonomická (Pejšová, 2008).

3.1. Struktura a rozhraní systému

NUŠL se skládá ze dvou hlavních komponent: vyhledávacího systému využívajícího software FAST ESP a vlastního repozitáře NUŠL fungujícího na softwaru Invenio.

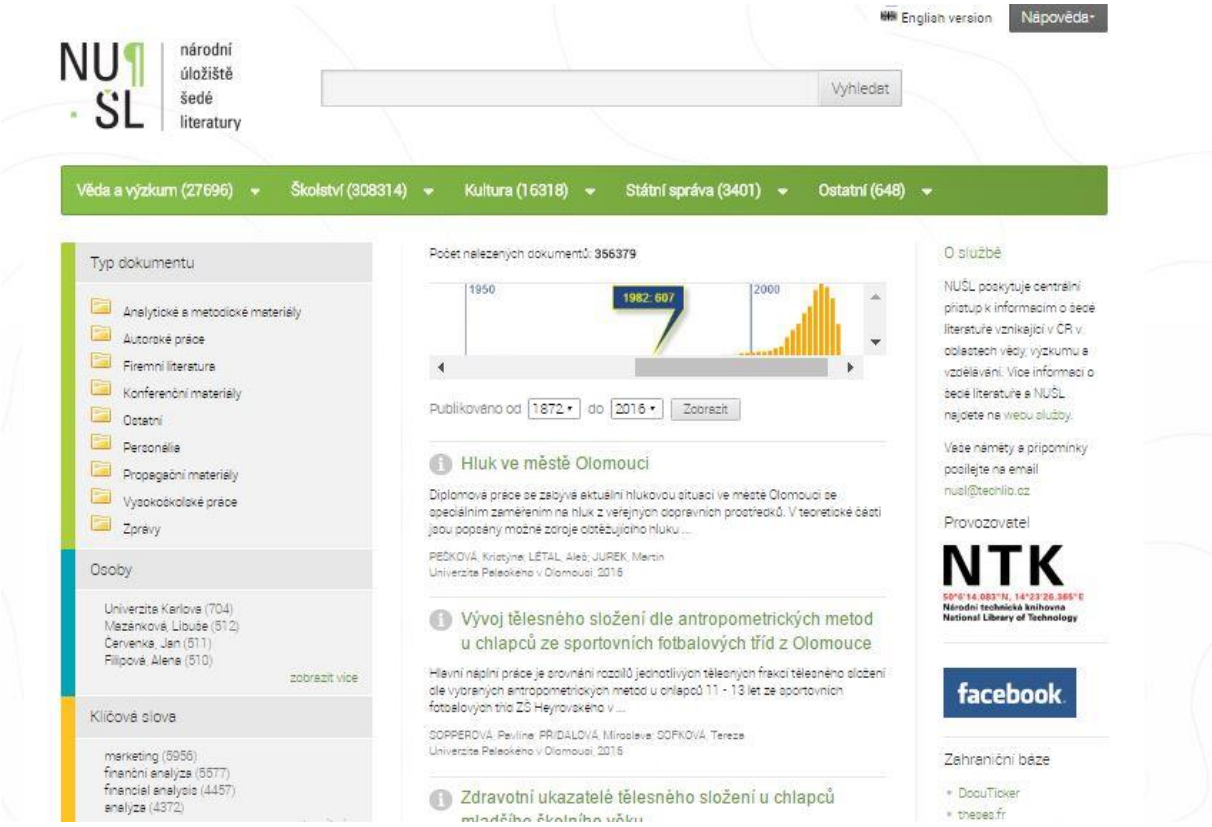
Vlastní repozitář NUŠL v Inveniu poskytuje agregační i archivační funkci. Funguje jako agregátor bibliografických záznamů, které jsou získávány z jiných systémů pomocí protokolu OAI-PMH či pravidelným dávkovým dodáváním, a zároveň slouží jako přímé úložiště záznamů a dokumentů (vkladatelé vytvářejí záznamy přímo v systému a případně přiloží i soubor dokumentu).

Všechny záznamy v NUŠL Invenio (bez ohledu na jejich zdroj) jsou zobrazitelné a adresovatelné. Zároveň je možné k nim v tomto systému přidávat plné texty či je jinak upravovat, a to hromadně i individuálně. Data z tohoto systému jsou dále stahovaná pomocí protokolu OAI-PMH do dalších systémů (například OpenAIRE, BASE atd.). Obrázek 7 ukazuje hlavní stránku uživatelského rozhraní repozitáře NUŠL Invenio.

[12] Na počátku projektu se ještě jednalo o Státní technickou knihovnu, která se na Národní technickou knihovnu přejmenovala v roce 2009.

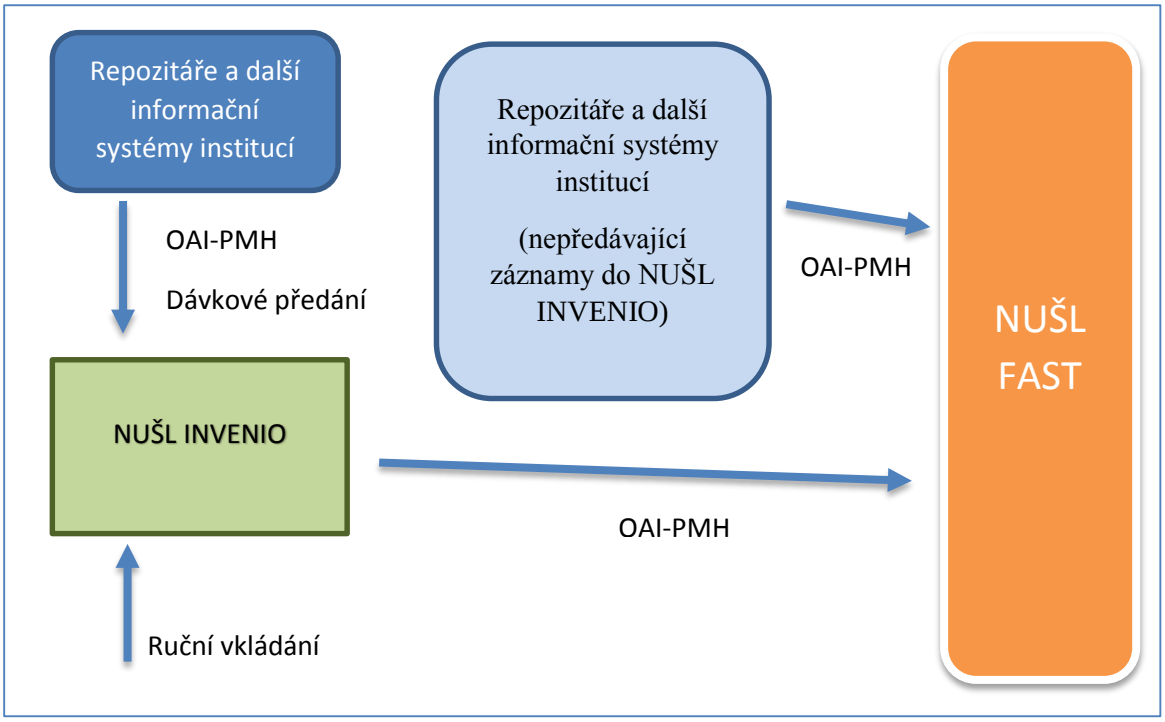
Obrázek 7: Úvodní stránka uživatelského rozhraní NUŠL Invenio (aktuální ke 12. 5. 2016)

Vyhledávací systém FAST pak tvoří další vrstvu nad systémem NUŠL Invenio (hlavní stránka uživatelského rozhraní systému NUŠL FAST je zobrazena na obr. 8).



Obrázek 8: Úvodní stránka vyhledávacího rozhraní NUŠL FAST (aktuální ke 12. 5. 2016)

Vyhledávací systém FAST vyhledává nad indexem složeným z dat z repozitáře NUŠL Invenio a dalšími systémy zapojenými do partnerské sítě NUŠL, jejichž data nejsou na základě smlouvy stahována do systému NUŠL Invenio (na obr 9. je schéma znázorňující popsanou strukturu).



Obrázek 9: Schéma struktury NUŠL

Z jednotlivých zdrojů jsou data sklízena přes OAI-PMH protokol, indexována a následně zpřístupňována v uživatelském rozhraní systému FAST, které umožňuje využívat nadstavbové vyhledávací funkce (filtrování výsledků podle typu dokumentu, autora, klíčových slov, jazyka dokumentu nebo dostupnosti plného textu, omezení výsledků podle organizace nebo podle data publikování). Na rozdíl od systému NUŠL Invenio není možné v systému NUŠL FAST odkazovat na konkrétní záznamy, ale pouze na stránky s výsledky s konkrétně nastavenými parametry (ovšem výsledky se v čase mění). Není ani možné záznamy upravovat. Změny je možné provádět pouze úpravou konvertoru, který převádí záznamy převzaté z jiného systému, a následně novým sklizením dat. Jedním z hlavních důvodů implementace další vrstvy nad systémem Invenio bylo, že ne všechny spolupracující instituce souhlasily s uložením svých záznamů v Inveniu, nicméně souhlasily s využitím dat ve vyhledávacím systému NUŠL FAST (právě kvůli jeho omezením).

Z popisu této architektury vyplývá, že přestože lze všechny záznamy z NUŠL Invenia nalézt pomocí systému FAST, ne všechny záznamy z NUŠL FAST lze najít v NUŠL Invenio. I přesto

jsou záznamy uložené v systému Invenio vhodnější pro testování další práce s daty, protože je možné do nich data snadno přidávat či s nimi dále různě pracovat.

3.2. Rozsah systému NUŠL

Jak napovídá již název samotný, Národní úložiště šedé literatury se zaměřuje na shromažďování šedé literatury na národní úrovni. Šedá literatura je již ze své samé podstaty rozmanitá, jak dokládá i jedna z nejužívanějších definic šedé literatury[13], která ji vymezuje jako „produkovanou na všech úrovních vládních, akademických, obchodních a průmyslových institucí jak v elektronické, tak v tištěné podobě, které neprošly standardním vydavatelským procesem či nejsou distribuovány do standardní prodejní sítě, tj. vydávány institucemi, jejichž hlavní činností není vydavatelská činnost[14]“ (Farace, 2010, s. 7). Pražská definice[15] šedé literatury pak kromě výše zmíněného uvádí, že do šedé literatury patří dokumenty „dostatečně kvalitní na to, aby byly shromažďovány a uchovávány v knihovnách a institucionálních repozitářích” (Schöpfel, 2010).

Dle těchto definic není šedá literatura nijak omezena podle typu dokumentu či jeho obsahu, ale je vymezena pouze producentem dokumentu a způsobem jeho zveřejnění, případně určitou mírou kvality (byť toto vymezení je velmi vágní a problematické).

Již v počátku projektu NUŠL musela být tedy stanovena vlastní konkrétní pravidla určující, jaké dokumenty do systému patří.

Geografické hledisko vymezuje místo vzniku dokumentů na území České republiky bez ohledu na jazyk dokumentu.

Původně bylo stanoveno i časové omezení, které umožňovalo vkládat pouze dokumenty vzniklé od roku 2009, nicméně od tohoto limitu se v pozdějších letech upustilo a nyní jsou v systému NUŠL i záznamy dokumentů mnohem starších. V systému NUŠL Invenio je nejstarší záznam dokumentu z roku 1920 a ve vyhledávacím rozhraní je možné najít i záznam dokumentu z konce 19. století.

[13] Označována jako Lucemburská definice, protože byla oficiálně představena na 3. ročníku Mezinárodní konferenci o šedé literatuře konajícím se v roce 1997 v Lucembursku.
[14] Překlad podle Pejšová, 2010.
[15] Představená v roce 2010 v Praze během 12. ročníku Mezinárodní konference o šedé literatuře.

V otázce obsahu dokumentů je záběr systému vymezen široce na oblast vědy, výzkumu, vzdělávání a kultury. Zároveň je snaha o to, aby vkládané dokumenty obsahovaly informace, které nemají pouze krátkodobý význam (pozvánky na akce apod.).

Dokumenty, jejichž záznamy jsou do NUŠL vkládány, musí také odpovídat dané typologii dokumentů. NUŠL používá s několika obměnami typologii, která byla vytvořena na počátku projektu. Současná podoba typologie dělí dokumenty do šesti hlavních skupin, v nichž jsou pak už zařazeny jednotlivé typy dokumentů (každý záznam je zařazen právě do jednoho z typů dokumentů):

1. Zprávy (cestovní, grantové, statistické, výzkumné, z projektu, výroční…)
2. Autorské práce (monografie, tematické sborníky, preprinty, referáty)
3. Firemní literatura (firemní tisk, katalogy výrobků, věstníky)
4. Konferenční materiály (programy, sborníky, příspěvky, postery)
5. Analytické a metodické materiály (analýzy metodiky, studie)
6. Propagační a vzdělávací materiály (brožury, letáky, katalogy výstav…)
7. Vysokoškolské kvalifikační práce (bakalářské, diplomové, disertační, rigorózní a habilitační práce)

3.3. Zdroje záznamů

Jak vyplývá z definice šedé literatury, producent dokumentu je pro šedou literaturu jedním ze zásadních činitelů. NUŠL nemá nastavena žádná speciální pravidla pro výběr institucí (či jiných partnerů) ke spolupráci. Jediným předpokladem pro možnost zahájení spolupráce s NUŠL je produkce šedé literatury odpovídající výše uvedeným pravidlům.

Rozsah institucí produkujících dokumenty, od nichž jsou záznamy (a případně i dokumenty) přebírány do NUŠL, odpovídá zmiňovaným definicím a zahrnuje tedy vládní, akademické, obchodní i průmyslové instituce a ve výjimečných případech dokonce i fyzické osoby.

V současné době používá NUŠL dělení zdrojových institucí do těchto 5 hlavních kategorií, které již mohou obsahovat jak další podkategorie, tak v některých případech přímo již konkrétní instituce:

- Školství (veřejné a soukromé vysoké školy),
- Kultura (galerie, muzea, knihovny a do této kategorie spadá i Národní památkový ústav)

- Věda (ústavy Akademie věd ČR, vědecké výzkumné instituce v.v.i. a další výzkumné ústavy)
- Státní správa (ministerstva a další instituce jako Parlamentní institut, Státní úřad pro jadernou bezpečnost, Státní zemědělská a potravinářská inspekce a Úřad průmyslového vlastnictví)
- Ostatní (momentálně obsahuje Českou národní banku, řadu neziskových organizací a dva osobní archivy).

Kromě typu se od sebe instituce výrazně liší i oborem zájmu, a to jak v rámci výše zmíněných kategorií (různé výzkumné ústavy se zabývají velice rozdílnými obory), tak uvnitř jednotlivých institucí (v rámci jedné vysoké školy mohou fungovat součásti zabývající se technickými, sociálními i humanitními vědami).

3.4. Způsoby spolupráce a získávání záznamů

Způsoby spolupráce jdou ruku v ruce se způsobem získávání záznamů. Pro přidání záznamů do NUŠL je nutné uzavřít licenční smlouvu s institucí (nebo osobou), která vykonává majetková autorská práva příslušných děl. Nejčastěji se jedná právě o instituce vykonávající práva k zaměstnaneckým a školním dílům. S fyzickými osobami se smlouvy uzavírají pouze v případě tzv. osobních archivů, ale tato praxe není zatím příliš běžná.

Pro uzavření spolupráce jsou zásadní dvě rozhodnutí, a to rozsah předávaných dat a způsob jejich předávání.

Otázka rozsahu dat se netýká pouze výběru poskytovaných záznamů (omezení na typ dokumentu, časové období apod.), ale i rozhodnutí, zda budou předávány pouze záznamy, nebo i plné texty. V případě neposkytování plných textů přímo do systému NUŠL Invenio je nutné předávat buď odkaz do nějakého vlastního systému instituce, či údaj značící, kde by byl dokument v případě zájmu dostupný.

Způsob předávání dat je v zásadě dvojí: předávka dat z jiného systému a ruční vkládání do Invenia. Záznamy mohou být předávány z jakéhokoliv jiného systému, který umožní export dat ve strojově čitelném formátu. Nemusí se tedy jednat přímo o repozitář, ale může jít o studijní informační systém, knihovní katalog či jiný informační systém. Pokud to příslušný systém zdrojové instituce umožňuje, jsou data stahována automatizovaně pomocí protokolu OAI-PMH. Není-li to možné, je při sjednávání spolupráce smluvené dávkové předávání. Takto

získaná data jsou v závislosti na smlouvě předávána buď do repozitáře NUŠL Invenio, nebo pouze do indexu vyhledávacího rozhraní NUŠL.

Druhou možností předávání dat do sytému NUŠL je pak jejich přímé vkládání do NUŠL Invenio. Určený pracovník partnerské instituce získá přístupové údaje do systému NUŠL Invenio a je buď proškolen, nebo alespoň obdrží sadu návodů pro vkládání. Vlastní vkládání probíhá ve webovém rozhraní systému Invenia pomocí jednoduchého formuláře. Po vložení několika prvních záznamů je vkladatelům poskytována zpětná vazba, aby došlo k objasnění problémů a alespoň elementárnímu sjednocení přístupů ke vkládání.

3.5. Metadatové záznamy v systému NUŠL Invenio

Metadatové záznamy musí být dle platných Pokynů pro zpracování záznamů v systému Invenio vytvořeny v českém nebo anglickém jazyce a v případě dokumentů v jiných abecedách je nutné identifikační údaje transkribovat do latinky (Frantíková, 2014).

Interním formátem softwaru Invenio je MARC 21, takže záznamy jsou uloženy v tomto formátu a do ostatních metadatových formátů jsou druhotně převáděny.

Je dán seznam povinných polí, která musí být vyplněna ve všech záznamech. U části z nich je obsah generován systémem, ostatní je třeba buď vyplnit, nebo identifikovat v přebíraných datech. Při ručním vkládaní systém zabrání uložení záznamu, pokud nejsou vyplněna všechna povinná pole, a nahlásí vkladateli problém. Některá povinná pole jsou vázána na konkrétní typ dokumentů a u jiných typů se vůbec nevyskytují.

3.5.1. Seznam polí povinných pro všechny typy dokumentů:

Následuje přehled polí (názvy odpovídají pojmenování v systému) a slouží pro ilustraci podoby záznamu v NUŠL.

- Identifikátor – jedinečný identifikátor je přiřazován automaticky systémem Invenio a tvoří základní součást URI.
- Název
- Název v angličtině – Tento údaj je sice uváděn mezi povinnými poli, nicméně existují i záznamy bez tohoto názvu a i při vytváření záznamu přímo v Inveniu není nevyplnění tohoto pole překážkou pro dokončení záznamu.
- Autor – Údaj je povinně zapisován v jednotném tvaru (příjmení, jméno) bez akademických i jiných titulů. Počet autorů není omezen. Autorem může být fyzická i právnická osoba i kombinace obojího.

- Datum zveřejnění zdroje – datum je povinně zadáváno v jednotném tvaru (RRRR-MM-DD a odvozených kratších tvarech RRRR-MM a RRRR)
- Předmět – heslo PSH. Tento údaj nelze vynucovat u hromadného předávání záznamů. Vkladatelé vkládající přímo do Invenia NUŠL jsou instruováni je vkládat, nicméně tato povinnost není softwarově ošetřena. Chybějící hesla PSH jsou pak řešena pomocí automatické indexace.
- NUŠL typ dokumentu – při přímém vkládání je typ dokumentu určen již na začátku procesu a na základě tohoto výběru se vkladateli zobrazí šablona. U hromadné předávky dat musí být nastavena převodní tabulka.
- Jazyk dokumentu – jazyk je označen tříznakovým kódem podle normy ISO 639-2 Kódy pro názvy jazyků - Část 2: Třípísmenný kód.
- Autorská práva – automaticky je doplňován text „Dílo je chráněno podle autorského zákona č. 121/2000 Sb." s odkazem na text příslušného zákona. Pokud je dokument zveřejňován pod veřejnou licencí Creative Commons, je možné v tomto poli vybrat z roletky příslušnou variantu a automaticky bude doplněn i text na znění licence.
- Název instituce
- Kontaktní informace - vyplňovány automaticky systémem.

Povinná pole pro záznamy konferenčních materiálů:

- Název konference/akce
- Místo konání konference/akce
- Datum nebo rozmezí konání konference/akce

Povinná pole pro záznamy vysokoškolských kvalifikačních prací:

- Akademický titul
- Typ studia
- Studijní obor
- Instituce přidělující titul
- Datum obhajoby

Další pole jsou již nepovinná a jejich vyplnění závisí na vkladateli nebo dostupnosti daných informací ve zdrojovém systému. Doporučováno je vkládání a přebírání údajů vždy, když jsou k dispozici.

Nepovinná pole společná pro záznamy všech typů dokumentů:

- Podnázev v jazyce dokumentu
- Podnázev v angličtině
- Název části dokumentu
- Číslo části dokumentu
- Přispěvatel - další autorské údaje
- Nakladatel
- Datum změny zdroje
- Poznámka
- Předmět – volně tvořená klíčová slova
- Abstrakt
- Bibliografická citace
- Autorská práva – CC licence
- Rozsah dokumentu
- Poznámka k souboru
- Název edice/série
- Číslo svazku edice/série
- Dostupnost - automaticky generovaný údaj, který se vyskytuje u záznamů bez přiloženého dokumentu.

Nepovinná pole specifická pro typ dokumentu „zprávy":

- Identifikační číslo projektu nebo kód vědeckovýzkumné práce
- Poskytovatel projektu

Nepovinná pole specifická pro typ dokumentu „konferenční materiály":

- Varianta názvu konference nebo jiné akce v cizím jazyce

Nepovinná pole specifická pro typ dokumentu „konferenční materiály", „autorské práce" a „firemní literatura":

- ISBN
- ISSN
- Název zdrojového dokumentu
- ISBN zdrojového dokumentu
- ISSN zdrojového dokumentu
- Informace o propojení

Záznamy je možné získat stažením pomocí protokolu OAI-PMH ve formátech OAI DC, MARCXML a NUŠL nebo zobrazit jednotlivě v systému Invenio ve formátech MarcXML, Dublin core, NUŠL a RIS.

3.5.2. Věcný popis

Specifická pozornost v této práci je věnována věcnému popisu v rámci systému NUŠL Invenio. V současné době se v repozitáři NUŠL nacházejí tři skupiny polí věcného popisu.

Jedná se o volná klíčová slova, která jsou v záznamech umístěná v poli 653 0_. Do této nejvolnější kategorie spadají samozřejmě volná klíčová slova, která takto vloží vkladatelé přímo do Invenia, ale především veškerá klíčová slova získaná automatizovaně z jiných systémů. Jiné systémy sice mohou používat vlastní řízené slovníky či jiné klasifikační systémy, ale v systému NUŠL se dostanou do nejobecnější kategorie klíčových slov. V podpolích jsou sice uchovávány údaje o původním slovníku a případně i identifikátor, ale tyto údaje nejsou zveřejňovány v běžném zobrazení a bez dalšího zpracování ani není způsob, jak jich využít.

Další dvě skupiny již tvoří hesla PSH. Jedná se o hesla PSH přidělená při ručním vkládání záznamu do systému NUŠL a o hesla PSH přiřazená pomocí automatické indexace. Ručně přiřazená hesla jsou uložena v poli 650 _7, zatímco hesla získaná automaticky jsou v poli 650 27. Automaticky přidělená hesla nejsou v záznamu při běžném prohlížení na webu vidět, ale slouží hlavně k dalšímu strojovému zpracování pro předávání do dalších systémů (viz kapitola 5.2.).

4. Polytematický strukturovaný heslář

Polytematický strukturovaný heslář (dále též jako PSH) je v systému NUŠL používán jak k intelektuální, tak k automatické indexaci, a proto bude v dalších částech této práce vybrán jako cílový slovník prováděného mapování. Následuje představení jeho struktury a užití, které hrají zásadní roli při propojování s dalšími pořádacími systémy.

4.1. Základní přehled

PSH je věcný selekční jazyk vyvíjený a spravovaný od počátku 90. let Národní technickou knihovnou. V NTK i dalších institucích je využíván k indexaci knihovního fondu i k dalším činnostem. Používán je i k indexaci dokumentů (manuální i automatické) v Národním úložišti šedé literatury.

Jedná se o selekční jazyk se stromovou strukturou, který má 44 nejvyšších bodů (tzv. hlavní hesla) reprezentujících tematické řady, pod něž jsou pak v hierarchické struktuře podřazena všechna ostatní hesla. Každé heslo může mít kromě preferovaného znění i znění nepreferované a může být propojeno s jiným heslem pomocí odkazu „viz též“. Jednou z hlavních vlastností PSH je, že každé heslo se může ve struktuře objevit pouze jednou a může mít pouze jedno nadřazené heslo.

4.2. Historie a užití PSH

PSH vznikal ve Státní technické knihovně již od počátku 90. let, kdy byl jeho vývoj podpořen několika projekty[16]. Postupně vznikaly verze 0 (1994) a verze 1.0 (1996). Od roku 1995 jsou všechny bibliografické záznamy dokumentů v STK při katalogizaci označovány i hesly PSH (Smolka, 1998).

V roce 1997 začala distribuce PSH dalším knihovnám i jiným zájemcům, tato služba byla zpoplatněna až do roku 2009, kdy byl heslář licencován pod veřejnou licencí Creative Commons (Kožuchová, 2010).

V NTK je PSH dodnes využíván ke katalogizaci knih a článků časopisů a kromě toho i k popisu dokumentů v Institucionálním repozitáři NTK a repozitáři NUŠL. Mimo NTK

[16] 1991 – 1993: I 096 Technologie poloautomatické indexace dokumentů s použitím selekčních jazyků verbálního typu pro polytematické fondy.
1995 – 1996: RS95IF074 Zajištění dostupnosti informací v knihovnách z hlediska věcného přístupu prostřednictvím polytematického strukturovaného slovníku (tezauru).

používají PSH pro katalogizaci knihovny Českého vysokého učení technického v Praze, Ústřední knihovna Vysokého učení technického v Brně, Vědecká knihovna v Olomouci, knihovna Západočeského muzea v Plzni či Ústřední knihovna Filozoficko-přírodovědecké fakulty Slezské univerzity v Opavě (Národní technická knihovna, 2015). Kromě tohoto klasického využití je PSH implementován i ve webové aplikaci určené pro vyhledávání zaměstnanců Univerzity Pardubice, kde slouží k vyhledávání podle vědecko-výzkumného zaměření (Univerzita Pardubice, 2015).

4.3. PSH jako selekční jazyk

PSH je definován jako selekční jazyk, tedy „umělý informační jazyk používaný k vyjádření identifikačních nebo obsahových selekčních údajů za účelem pořádání, ukládání a vyhledávání dokumentů“(Balíková, 2003a). Selekční jazyky se dělí podle několika hledisek:

1) Typ zpřístupňovaných údajů
   a) Dokumentační
   b) Faktografické
2) Charakter zpřístupňovaných údajů
   a) Identifikační
   b) Věcné
3) Povaha komplexních pořádacích znaků
   a) Prekoordinované
   b) Postkoordinované

PSH byl primárně vytvářen k indexaci knihovního fondu, jedná se tedy o dokumentační selekční jazyk, „určený k pořádání a ukládání dokumentografických informací, tj. selekční jazyk zachycující údaje, které poskytují formální a obsahovou charakteristiku dokumentu“ (Balíková, 2003b).

Z hlediska charakteru zpřístupňovaných údajů se jedná o věcný selekční jazyk, který je v TDKIV definován jako:

„Selekční jazyk používaný pro zpracování dokumentů pomocí věcných údajů s cílem umožnit vyhledávání dokumentů podle obsahu. Podle typu používaných lexikálních jednotek se vyčleňují věcné selekční jazyky na bázi přirozeného jazyka označované jako předmětové selekční jazyky a věcné selekční jazyky umělé, které se označují jako systematické selekční jazyky. Podle způsobu organizace lexikálních jednotek v procesu ukládání a vyhledávání, se vyčleňují prekoordinované a postkoordinované selekční jazyky“ (Balíková, 2003c).

V otázce, zda se jedná o postkoordinovaný, nebo prekoordinovaný jazyk, se odborníci zcela neshodují. Na tento problém ve své diplomové práci upozornila Linda Skolková (2007a, s. 1), která se spolu s dalšími autory „přiklání spíše k zařazení hesláře mezi systémy postkoordinované“. Uznává ovšem, že v PSH se uplatňuje i princip prekoordinace, a není tedy možné jej jednoznačně zařadit.

U věcných selekčních jazyků je dále možné sledovat tyto atributy (Kučerová, 2005):

1) Oborový záběr
   a) Univerzální jazyky (polytematické)
   b) Speciální jazyky (oborové)
2) Jazyk
   a) Jednojazyčné selekční jazyky
   b) Vícejazyčné selekční jazyky
3) Postup při konstrukci
   a) Top-down
   b) Bottom-up
4) Míra granularity (hrubosti klasifikace)

Jak již jeho název napovídá, PSH je univerzálním věcným selekčním jazykem, který se snaží postihnout „všechny základní oblasti lidského poznání“ (Skolková, 2007b, s. 4), což se odráží i v jeho 44 hlavních kategoriích.

Jelikož PSH obsahuje deskriptory a nedeskriptory v české a anglické jazykové verzi, jedná se o vícejazyčný selekční jazyk. Podle Jánské je „u vícejazyčných tezaurů důležitý vzájemný poměr jazyků, tj. jejich status. Je důležité ustanovit určitý jazyk jako výchozí či zprostředkující, dominantní nebo sekundární. Ten jazyk, jehož lexikální jednotka vyvolává zvláštní překladové problémy, je většinou označován jako výchozí jazyk. Je to jazyk, který je východiskem pro překlad deskriptoru do nejblíže ekvivalentní lexikální jednotky (nebo jednotek) druhého jazyka, resp. jazyka překladu“ (Janská, 2008). Tento poměr lze stanovit i u PSH, byť se nejedná o tezaurus. Výchozím jazykem je v tomto případě jednoznačně čeština, byť jsou samozřejmě i případy, kdy se pro anglický termín jen těžko hledá český ekvivalent. V PSH jsou tak k nalezení i případy hesel mající i v české verzi anglické znění (např. API http://psh.ntkcz.cz/skos/PSH13931).

Při původní konstrukci, kdy byla jednou za dané období vydána nová verze PSH, se postupovalo metodou top-down, tedy „od celistvé abstraktní představy ke konkrétní implementaci“ (Píšková, 2012, s. 10). Po ukončení verzování se již dá mluvit o tom, že doplňování funguje metodou bottom-up, protože termíny jsou doplňovány na základě analýzy logů vyhledávání a podnětů indexátorů průběžně v procesu používání. Vyvíjí se tedy „od jednotlivých lexikálních jednotek ke kompletnímu slovníku“ (Píšková, 2012, s. 10).

PSH má poměrně hrubou granularitu, a proto bývá používán společně s dalšími selekčními jazyky. Tato hrubost je zakotvena v Pravidlech pro správu a aktualizaci, která upozorňují, že v PSH nemají místo příliš podrobné termíny, a i proto je dán limit šesti, výjimečně sedmi úrovní hloubky.

Konečné určení PSH jako konkrétního druhu věcného selekčního jazyka není jednoznačnou záležitostí. Přestože se v názvu definuje jako heslář, přikláněla se Kučerová (2005) ve své analýze spíše k zařazení mezi klasifikace nebo tezaury. Pro tuto práci není ale toto zařazení zásadní a jde spíše o utvoření představy o PSH, a proto pokud se budeme držet pojmu heslář, jde spíše o označení PSH coby konkrétního produktu než předmětového hesláře jako typu věcného selekčního jazyka.

4.3.1. Vztahy v PSH

Pravidla pro syntagmatické vztahy mezi termíny, které jsou dány kontextem při indexaci dokumentu nebo jeho vyhledávání, nejsou v PSH nijak stanoveny a odtud plynou i nejasnosti ohledně jeho postkoordinačního/prekoordinačního charakteru. Pouze ve starší verzi Indexačních pravidel pro práci s PSH je uvedeno doporučení využít prekoordinovaný termín, pokud je to možné. Pokud takový termín heslář neobsahuje, je na místě využít postkoordinaci, ale pouze za předpokladu, „že při užití kombinace hesel nedojde ke ztrátě původního významu“ (Skolková, 2007b, s. 12).

Paradigmatický vztah, tedy vztah „mezi pojmy, popř. výrazy, který existuje nezávisle na větném kontextu“ (Hrazdil, 2003), je v PSH trojího druhu: hierarchie, ekvivalence a asociace.

Hierarchický vztah, který TDKIV definuje jako „formální vztah mezi dvěma entitami (termíny, třídami), kde jedna je podřízena druhé“ (Balíková, 2003d), pozorujeme mezi nadřazenými a podřazenými hesly. PSH má stromovou strukturu, přičemž je vlastně rozdělen

podle 44 „hlavních hesel“ do tohoto počtu menších stromů, pod něž jsou podřazeny ostatní termíny. V PSH je stanoveno, že každý termín se může ve struktuře vyskytovat pouze jednou, a proto má vždy pouze jedno nadřazené heslo. Toto omezení je pak nutno kompenzovat využitím asociačních vztahů k naznačení propojení mezi jednotlivými podstromy.

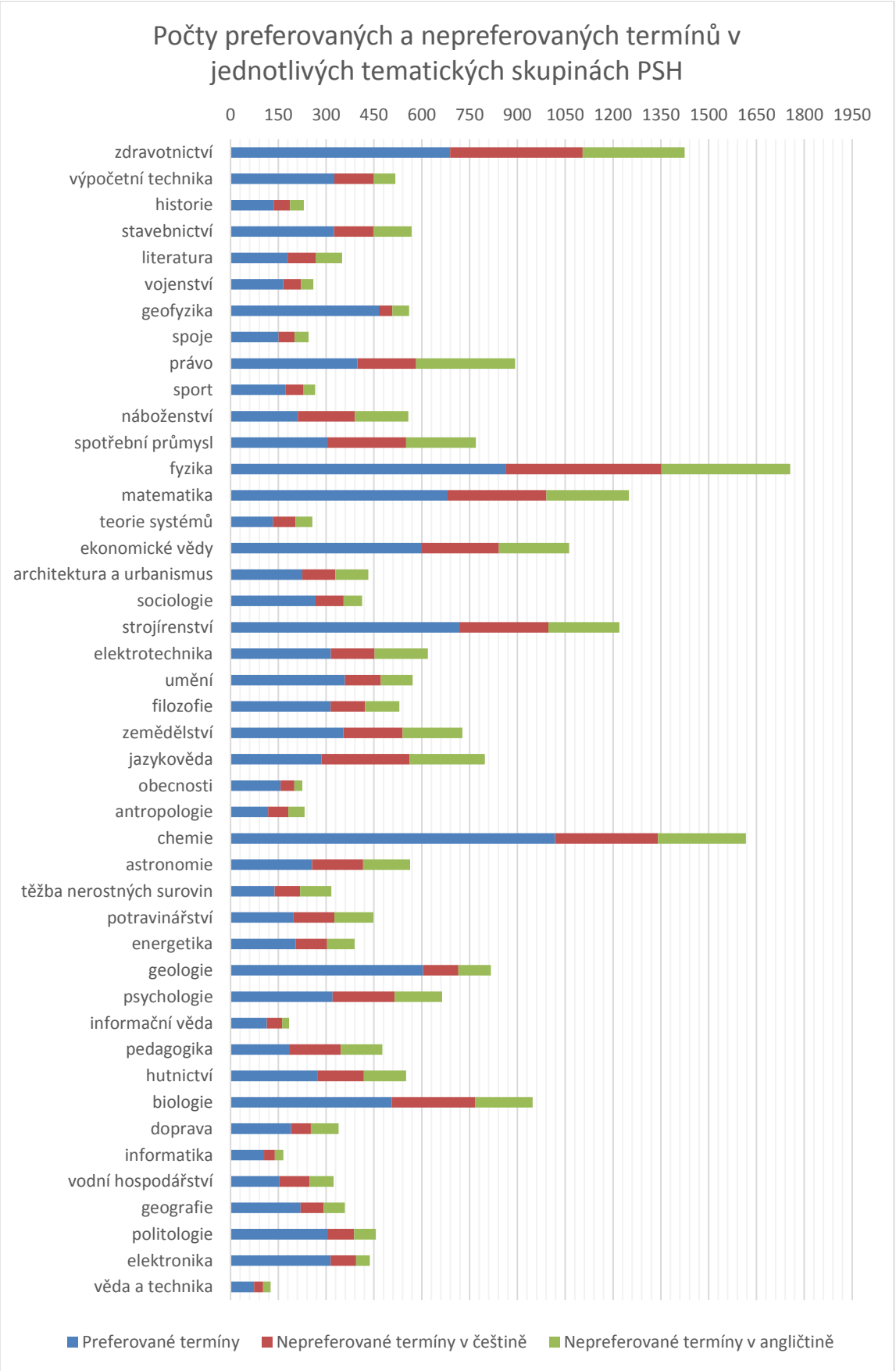
Ekvivalenční vztahy jsou v PSH mezi deskriptory (preferovanými termíny) a nedeskriptory (nepreferovanými termíny). „V řadě případů je kvůli přehlednosti hesláře jako celku využito hierarchizace vztahu ekvivalence, tj. některé termíny vyjadřující užší pojmy jsou zařazeny jako termíny nepreferované u hesel vyjadřujících odpovídající širší pojmy“ (Kožuchová, 2012, s. 4). Tento vztah je v rámci PSH vyjádřen jako vylučovací („viz odkaz“).

Asociativní vztah je v TDKIV definován jako: „reciproční sémantický vztah mezi lexikálními jednotkami, které nejsou členy stejné třídy ekvivalence a mezi kterými nelze definovat hierarchický vztah, avšak funkční, formální, prostorová nebo časová souvislost mezi nimi je natolik závažná, že je účelné pro potřeby vyhledávání jejich vzájemnou vazbu vyjádřit“ (Balíková, 2003i). Této definici odpovídá i použití v PSH, kde je používán často kvůli ukázání vzájemného propojení napříč jednotlivými tematickými stromy. Asociativní vztah je v PSH vyjádřen přidružovacím odkazem („viz též“).

4.3.2. Struktura

V současné době má PSH celkem 44 hlavních kategorií a v nich celkem přes 14 000 deskriptorů[17]. Nedeskriptorů je celkem 11 885 (6 329 českých a 5 556 anglických). Rozložení těchto preferovaných a nepreferovaných termínů do hlavních kategorií je znázorněno v grafu na obr. 10.

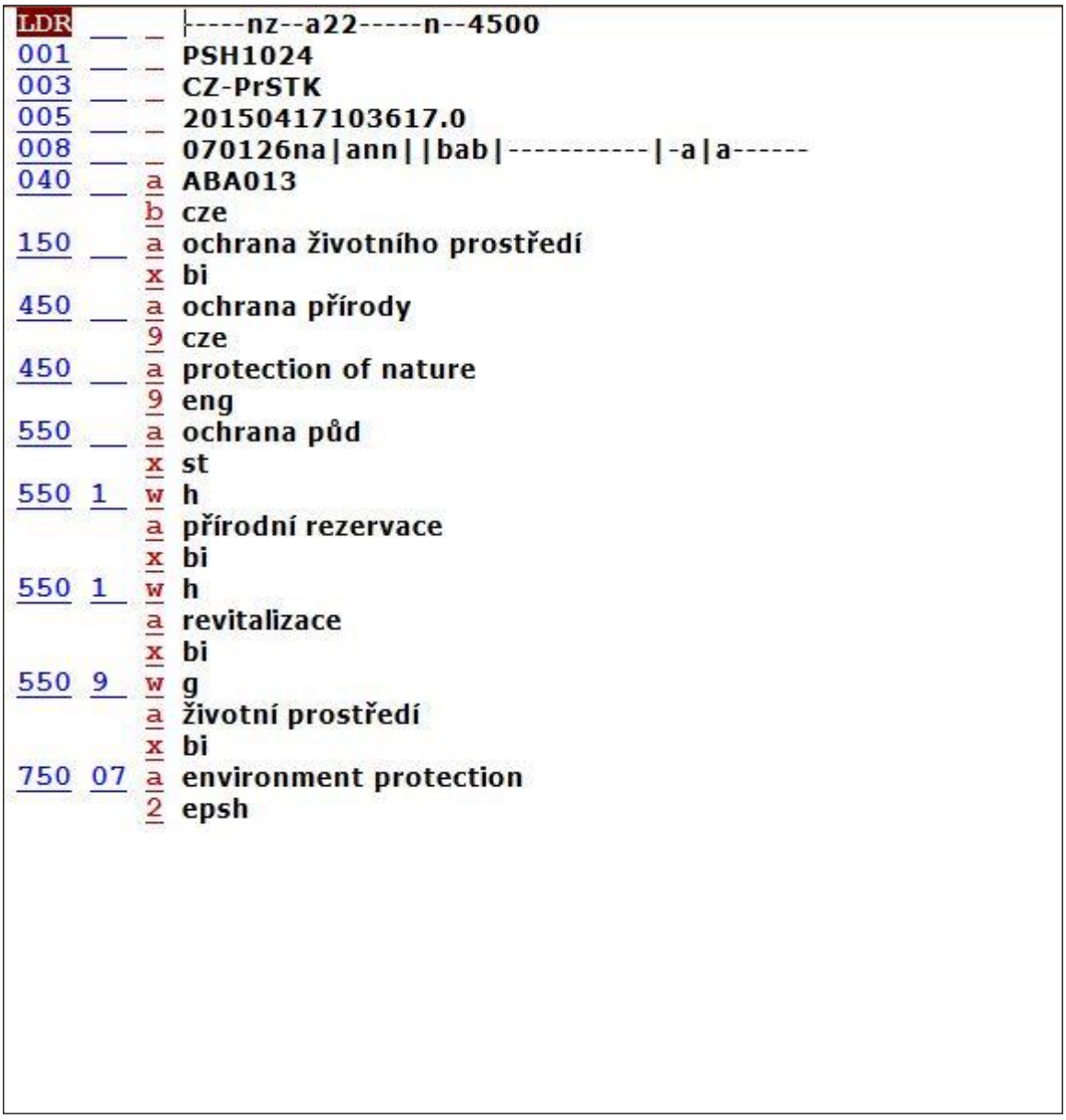
[17] V dubnu 2016 se jednalo přesně o 14 088 hesel.



Obrázek 10: Počty preferovaných a nepreferovaných termínů v jednotlivých tematických skupinách PSH

4.3.3. Záznamy PSH a jejich zveřejnění

Záznamy jsou uloženy a spravovány[18] v automatizovaném knihovním systému Aleph Národní technické knihovny. Toto uložení umožňuje katalogizátorům přímou práci s aktuálními daty PSH. Záznamy jsou uloženy v autoritním formátu MARC 21. Na obrázku 11 je zobrazen příklad záznamu hesla PSH v systému Aleph.



Obrázek 11: Záznam hesla PSH v systému ALEPH (aktuální ke 12. 5. 2016)

V poli 001 je uveden identifikátor PSH ID, který se skládá z prefixu PSH a čísla záznamu (jedná se zároveň o přírůstkové číslo). Na základě tohoto údaje je generován jednoznačný

[18] Správu má na starosti k tomu určený pracovník NTK. V době psaní této práce byla správa PSH i obsahu NUŠL v rukou jednoho člověka (autorky této práce), což právě vedlo k možnosti zkoumání dalších možností využití PSH v NUŠL.

identifikátor URI ve formátu http://psh.ntkcz.cz/skos/PSH_ID (např.: http://psh.ntkcz.cz/skos/PSH1024).

V poli 150 je české preferované znění termínu (podpole „a“) a dvoupísmenný kód tematické větve (podpole „x“).

Anglické preferované znění je zapsáno v poli 750 07 (podpole „a“).

Záznam může obsahovat řadu vylučovacích odkazů v poli 450 __ a, přičemž údaj o tom, zda je konkrétní nedeskriptor česky nebo anglicky, se uvádí pomocí třípísmenného kódu v podpoli 9. Přidružovací odkazy jsou v poli 550 1_ a. Vždy se odkazuje na českou verzi a je povinné doplnit do podpole „x“ kód větve, v níž se nachází odkazovaný záznam. Vylučovací i přidružovací odkazy jsou obousměrné a vyplňují se v obou propojovaných záznamech.

Údaje o hierarchickém zařazení jsou v poli 550 s indikátory 1_ a 9_. Záznam může mít řadu podřazených záznamů a tedy i polí 550 1_. Pole spojující s podřazeným záznamem obsahuje podpole:

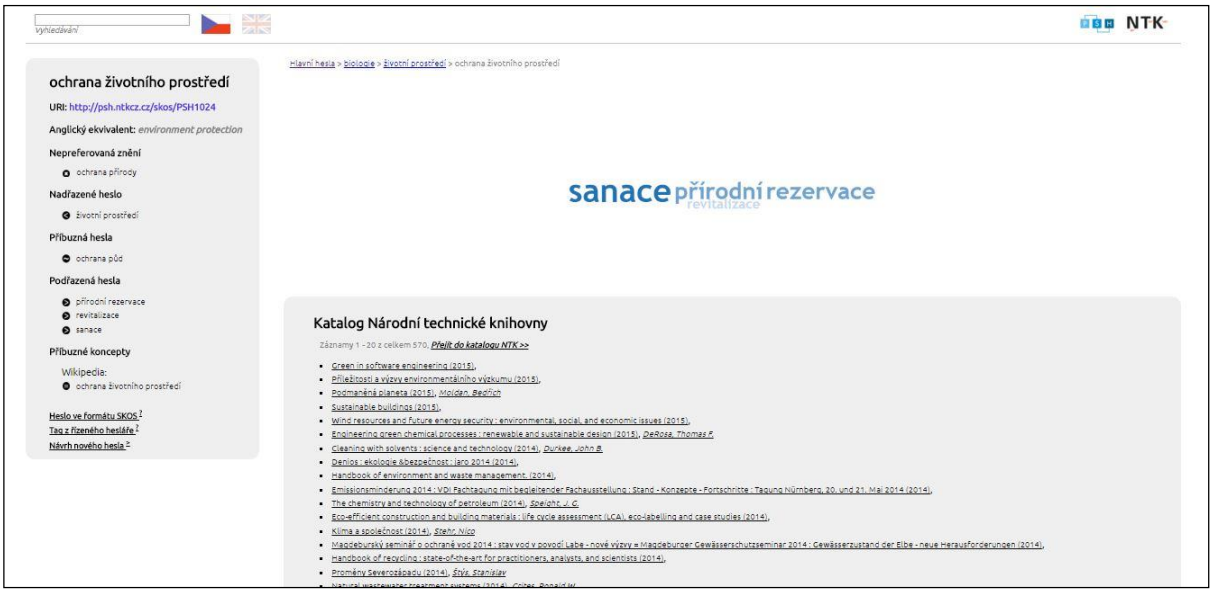
„w“, v němž je uveden kód „h“ signalizující, že se jedná odkaz na podřazené heslo,

„a“ obsahující vlastní znění podřazeného deskriptoru,

„x“ kód větve podřazeného deskriptoru (ta je vždy shodná s nadřazeným heslem).

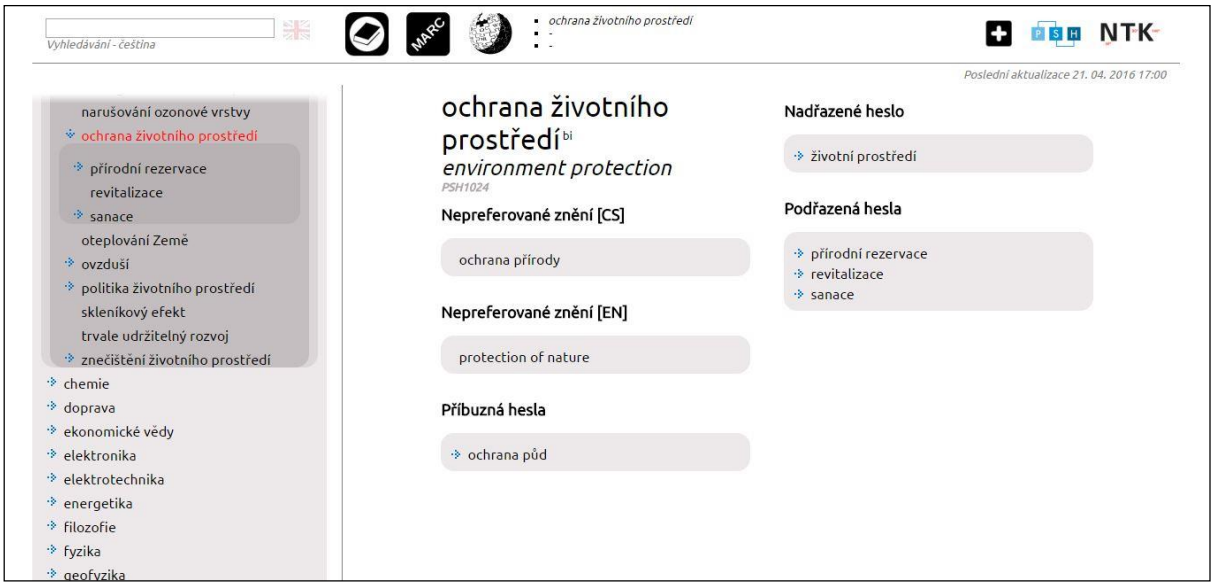
Pole 550 9_ označující nadřazené heslo může mít každý záznam pouze jedno. Podpole jsou obdobná jako u 550 1_, změna je pouze v podpoli „w“, kam se pro vyznačení odkazu na nadřazené heslo zapisuje kód „g“.

Z Alephu jsou pravidelně jednou za 24 hodin automaticky zasílány změny do externí databáze, která je základem pro webové aplikace PSH prohlížení a PSH manager.



Obrázek 12: Ukázka záznamu v rozhraní PSH prohlížení (aktuální ke 12. 5. 2016)

K prohlížení obsahu PSH je sice možné využít i Aleph NTK, ale primárně jsou k tomuto účelu určeny a používány tyto dvě aplikace. Data z Alephu jsou v externí databázi obohacena o odkazy na externí webové stránky (DBPedia a Wikipedia) a do katalogu NTK. Tyto údaje se zobrazují i ve zmiňovaných webových aplikacích. Zobrazení záznamu v aplikacích PSH prohlížení a PSH manager je ukázáno na obr. 12 a 13.



Obrázek 13: Ukázka záznamu v rozhraní PSH manager (aktuální ke 12. 5. 2016)

Data z externí databáze jsou převáděna do formátu SKOS v zápisu RDF+XML a takto jsou zveřejňována volně ke stažení na webových stránkách NTK pod veřejnou licencí CC BY-SA 3.0 - Uveďte autora-Zachovejte licenci 3.0 Česko.

5. Minulé snahy o sjednocování věcného popisu v systému NUŠL

Tato kapitola představuje minulé pokusy i dodnes používané metody pro sjednocování věcného popisu v NUŠL. Hlavním dříve rozvíjeným směrem byla automatická indexace, jejíž výsledky budou v této práci srovnávány s výsledky použití mapování. Kromě toho přináší kapitola i informace o současném využití tohoto sjednocení, které může v závěrečném pokusu sloužit i jako další možnost srovnání výsledků mapování a automatické indexace.

5.1. Automatická indexace v systému NUŠL

Automatickou indexací je v kontextu této práce myšleno přiřazování prvků selekčního jazyka (v NUŠL se jedná o hesla PSH) do záznamů. Vývoj automatické indexace v systému NUŠL probíhal obdobně jako v systému BASE popisovaném v kapitole 1.1. Stejně jako v případě německého akademického vyhledávače byla i v NUŠL nejdříve řešena automatizace indexace záznamů s plnými texty a následně i záznamů bez připojených plných textů dokumentů.

5.1.1. Automatická indexace s využitím plných textů

Řešení automatické indexace záznamů s plnými texty uloženými v NUŠL probíhalo v letech 2009–2011 a jako klasifikační schéma byl k indexaci vybrán Polytematický strukturovaný heslář (PSH). Oproti řešení použitému v BASE však v případě NUŠL nebyla snaha vytvořit zcela automatické řešení, které by jednoduše všem záznamům s plným textem uloženým v repozitáři přiřadil hesla, ale spíše nástroj, který měl vkladatelům do NUŠL (respektive obecně autorům šedé literatury) pomoci při výběru vhodných hesel PSH k popisu dokumentu.

Tento přístup vycházel z přesvědčení, že vzhledem k povaze šedé literatury, která není produkována a často ani dále zpracována klasickými způsoby, se do role indexátora dokumentů dostávají i neodborníci – ať už díla popisují v NUŠL nebo třeba na svých webových stránkách. Takový indexátor může být odborníkem na oblast své specializace, ale pravděpodobně nebude obeznámen s problematikou věcného popisu, konkrétním klasifikačním schématem používaným v daném systému a vztahem schématu k celému fondu. V případě systému, jako je NUŠL, je celá situace ještě umocněná množstvím osob vytvářejících záznamy, takže výsledný věcný popis dokumentů v systému trpí nekonzistencí a chybovostí (Mynarz, 2010).

Nástroj použitý v NUŠL měl pomoci s překonáním tohoto problému. Cílem bylo, aby na základě zpracování dokumentu pomocí automatických postupů nabídl vkladateli při vytváření záznamu vhodná hesla z řízeného slovníku PSH. Vkladatel by pak nemusel znát konkrétní řízený slovník, ani cokoli vědět o věcném pořádání.

Pro tento úkol byly testovány dva již existující nástroje, a to CDS Invenio BibClassify a Maui Indexer (Mynarz, 2009).

BibClassify je modul softwaru Invenio, který je v NUŠL používán pro vlastní repozitář dokumentů a záznamů. Nabízel se tedy jako vhodné řešení díky jednoduché implementaci modulu do zbytku NUŠL Invenia. Zmiňovaný modul je stejně jako zbytek Invenia šířen jako open source software, takže jeho použití nepřinášelo další náklady na nákup licence.

Tento nástroj provádí jednoduchou extrakci termínů cílového řízeného slovníku z textu daného dokumentu na základě frekvence výskytu. Modul umí pracovat s řízenými slovníky zapsanými ve formátu RDF/SKOS, a je tedy schopen pracovat s tím, že do počtu výskytů jednotlivých termínů započítává i výskyty jejich nepreferovaných znění. BibClassify nijak nevyužívá metod strojového učení ani umělé inteligence a veškeré činnosti provádí pouze jednoduchým porovnáváním pomocí regulárních výrazů (European Organization for Nuclear Research, 2008). Podle samotného manuálu modulu je výkon tohoto nástroje do značné míry závislý na kvalitě použitého řízeného slovníku a nejlepších výsledků dosahuje při použití bohatých a dobře strukturovaných tezaurů. Při testování použití tohoto nástroje v NUŠL se další slabinou ukázala jeho rychlost, která byla znatelně nízká obzvláště při pokusech o zpracování rozsáhlejších textů (Mynarz, 2009).

Druhým testovaným nástrojem, který byl nakonec vybrán k dalšímu použití, byl Maui Indexer[19]. Stejně jako BibClassify je i Maui Indexer šířen pod veřejnou licencí jako open source. Tento nástroj se také snaží indexovat dokumenty na základě automatické extrakce termínů určeného řízeného slovníku, ale v tomto případě k tomu používá „pokročilejší postupy, které zahrnují metody analýzy přirozeného jazyka a strojového učení“ (Mynarz, 2009).

Využití strojového učení při automatické indexaci vyžaduje nejprve vytvoření indexačního modelu (stejně jako v případě systému BASE). Tento model je „výsledkem strojového učení nad zdroji automatické indexace“ (Mynarz, 2011, s. 17), kterými jsou:

[19] Nástroj vyvinula Alyona Medelyan a momentálně je dostupný v softwarovém repozitáři na adrese: https://github.com/zelandiya/maui

1. Analýza indexovaného dokumentu
2. Analýza použitého řízeného slovníku
3. Analýza způsobu použití daného slovníku nad korpusem dokumentů

Model byl tedy vytvořen na základě množiny plných textů umístěných v NUŠL, které musely být intelektuálně indexovány pomocí cílového řízeného slovníku – v tomto případě PSH. Maui Indexer pak za použití metod strojového učení získá informaci o požadovaném způsobu používání hesláře při indexaci a následně je schopen po načtení neindexovaného předzpracovaného dokumentu navrhnout nejvhodnější termíny ze slovníku pro jeho popis. Předzpracování zahrnovalo převod do prostého textu, odstranění stop slov a stemming.

Jak bylo zmíněno, v případě NUŠL bylo cílem implementovat nástroj poskytující podporu vkladatelům záznamů do systému. Vkladatel po vytvoření identifikačního popisu dokumentu a nahrání plného textu nemusel sám vymýšlet vhodná klíčová slova a následně je hledat v PSH. Místo toho získal automatizované návrhy a pouze určil, zda dané termíny odpovídají obsahu dokumentu (obr. 14 zobrazuje obrazovku s automatickými návrhy hesel PSH). Hlavní rozhodování je tedy i nadále ponecháváno lidem, kteří by měli být schopni eliminovat chybně

Obrázek 14: Navrhované termíny z automatické indexace při popisu vkládaného dokumentu (Mynarz, 2011)

přidělená hesla, případně doplnit hesla chybějící. Díky využití nástroje schopného strojového učení je každá taková kontrola příspěvkem k vylepšování a zpřesňování systému, nikoli pouze kontrolou hesla mechanicky přiřazeného strojem.

Hlavním omezením tohoto nástroje je nutnost pracovat s plným textem dokumentu, který přímo v repozitáři NUŠL bývá velmi zřídka (přibližně 2 % záznamů mají přímo v Inveniu uložený plný text), a není tak možné ho použít na všechny záznamy v NUŠL. Od toho se pak odvíjí i další problém související s dalším strojovým učením indexačního modelu. Základní indexační model totiž vznikl na základě dokumentů vysokoškolských kvalifikačních prací z Vysoké školy ekonomické a dokumentů Akademie věd ČR, jejichž část byla pro tento účel označena a zaindexována hesly PSH. Model tak umí pracovat s dokumenty z některých oblastí (zejména ekonomie), ale stejným způsobem přistupuje i k dokumentům z naprosto odlišných oborů, což v důsledku způsobuje navrhování nevhodných hesel z nesprávných tematických větví PSH. V ideálním případě by model měl být schopný se tyto věci postupně naučit, pokud by měl k dispozici dostatek dokumentů označených PSH (nebo alespoň se zpětnou vazbou k návrhům automatické indexace). Tomuto procesu ovšem bránil zmiňovaný nedostatek plných textů přímo v repozitáři NUŠL a poměrně malé množství dokumentů vkládaných přímo ručně do NUŠL Invenia. Celou věc komplikuje i velký tematický záběr systému a široká typologie zpracovávaných dokumentů.

Pro případné zdokonalení nástroje by musely být vzaty v úvahu všechny tyto překážky a ty řešit například vytvořením více indexačních modelů podle oboru/instituce. Zároveň pořád platilo, že nástroj je možné použít pouze na záznamy s plnými texty, což by v případě NUŠL znamenalo něco přes 2 000 záznamů, tedy méně než 1,5 % z počtu záznamů v repozitáři Invenio.[20] V takovém případě by další vývoj znamenal spíše vyšší náklady než užitek, a tak se od tohoto projektu upustilo. V současné době se tedy tento nástroj v NUŠL nepoužívá, ale stále je k dispozici webová stránka http://invenio.ntkcz.cz/indexer/, na níž je možné vyzkoušet automatickou indexaci na libovolném českém textu.

5.1.2. Automatická indexace bez využití plných textů

Dalším pokusem ve sjednocování věcného popisu v systému NUŠL byl projekt „Automatická indexace obsahu Národního úložiště šedé literatury pomocí Polytematického strukturovaného hesláře“, jehož cílem bylo přiřadit každému záznamu, který ještě není indexován pomocí PSH, nejméně jedno odpovídající heslo z PSH.

[20] Stav k dubnu 2016.

Pro tento účel byla v roce 2012 vypsána veřejná zakázka, která poptávala indexaci předpokládaného počtu 130 000 záznamů pomocí PSH a vytvoření analýzy, dokumentace, uživatelského rozhraní pro kontrolu hesel a softwarového nástroje. Realizace tohoto projektu pak probíhala v roce 2013. Tuto zakázku získala firma Incad, která již měla zkušenosti s jiným projektem automatické indexace v této knihovně (v době, kdy ještě fungovala pod zkratkou STK), kdy dodávala software pro automatickou indexaci fondu knihovny pomocí LCC[21].

Už při přípravě zakázky počítal projekt s tím, že jedním z benefitů sjednoceného věcného popisu bude možnost generovat záznamům na základě hesel PSH i kategorie SIGLE, které jsou povinným prvkem popisu v databázi OpenGrey, kam prostřednictvím NUŠL přispívá NTK za celou Českou republiku.

Hlavními požadavky na software byla schopnost nástroje pracovat s údaji v českém jazyce, musel být open source a měl být napojitelný do klasifikačního modulu Invenio repozitáře NUŠL. Na základě těchto parametrů byl nakonec vybrán indexační a vyhledávací nástroj Lucene SOLR.

Za cílovou mapovanou strukturu byl již od počátku vybrán Polytematický strukturovaný heslář a jednalo se pouze o tom, do jaké hloubky stromové struktury se bude heslář využívat. Nakonec bylo rozhodnuto, že bude využíván celý strom PSH, který má 44 oborových větví a 6-7 úrovní hloubky.

Pro automatickou indexaci byly určeny všechny záznamy uložené v NUŠL Invenio, které neměly přiřazená hesla PSH ani manuálně (intelektuálně), ani pomocí předchozího projektu automatické indexace na základě plných textů. V průběhu projektu bylo v repozitáři Invenio 107 469 záznamů a z toho 24 587 již mělo nějakým způsobem přiřazené heslo PSH. Automatická indexace tedy měla do začátku za úkol zpracovat 82 882 záznamů (Kocourek, 2013). Tyto záznamy byly ve formátu MARCXML pomocí OAI-PMH protokolu načteny a zaindexovány do nástroje SOLR. Před samotnou automatickou indexací je nutné záznamy normalizovat, tedy odstranit redundantní informace a složité konstrukce. Normalizace v případě

[21] Při stěhování knihovny do nové budovy bylo rozhodnuto o tom, že volný výběr bude řazen podle systému LCC. Tyto údaje ovšem v záznamech nebyly a musely se ručně dodávat. Automatická indexace byla nasazena od roku 2008, protože při nejlepší vůli nebylo klasickými cestami možné včas označit a přestěhovat celý fond. Záznamy pak byly kontrolovány pracovníky STK, podle jejichž odhadů dosahovala automatické indexace „pouze“ 60 %. I tak dokázala automatická indexace 10krát zvýšit rychlost zpracování. Oproti tomu v případě NUŠL muselo být bráno v úvahu, že záznamy a k nim přidělené termíny PSH nebudou vždy jednotlivě kontrolovány a upravovány.

tohoto projektu znamená převedení vybraných polí formátu MARC záznamů do zjednodušeného indexu v nástroji SOLR. V následující tabulce 1 je zobrazeno, jaká pole jsou během procesu automatické indexace záznamů v NUŠL zpracovávána a jak jsou shlukována (INCAD, 2013).

	Označení pole ve formátu MARC 21	Označení v nástroji SOLR
Název	245a, 245b, 246a, 246b	Title
Anotace	520a	Annotace
Klíčova slova	650a, 650x, 653a	Subject
Konference	711g	Konference
Instituce	998a	Institution

Tabulka 1: Označení polí v MARC 21 a nástroji SOLR

Obdobným způsobem je třeba předem zpracovat i PSH. Heslář je načten do SOLR ve formátu SKOS a následně jsou zaindexována jednotlivá pole každého hesla: preferovaná znění hesla (v českém a anglickém jazyce), nepreferovaná znění hesla (v českém a anglickém jazyce), nadřazená, podřazená a příbuzná hesla, navigace stromem a hloubka umístění hesla ve struktuře PSH.

„Při přiřazování hesel je porovnáván normalizovaný zdrojový záznam s heslářem. Při porovnávání je hledána shoda na úrovni jednotlivých slov, či skupin slov“ (INCAD, 2013, s. 2). Hesla jsou vyhledávána v polích uvedených v tabulce: název, anotace, klíčová slova, konference, instituce. Pro hledání je využit i český gramatický slovník, který umožňuje vyhledávat i slova se stejným slovním základem.

Do systému bylo rovněž nutno implementovat slovník stop slov, tedy nevýznamových či v jazyce často používaných slov, která se nemají používat k automatické indexaci. Do tohoto slovníku musela být pro potřeby indexace NUŠL přidána i specifická hesla související s typologií repozitáře, jako je například „diplomová práce“ apod.

Míra shody je ohodnocena podle následujících bodů a váhy jednotlivých jejich parametrů. Pojmenování polí i parametrů je převzato z manuálu k aplikaci pro automatickou indexaci, a proto se vyskytuje ve zvláštní česko-anglické kombinaci (INCAD, 2013):

1) **Pole, ve kterém se dané slovo nachází ve zdrojovém dokumentu**. Hodnoty parametrů jsou vypočítávány tak, že se mezi sebou násobí váha daného pole, místo výskytu v heslář (pole a hloubka) a počet výskytů nalezeného slova v tomto konkrétním poli. Parametry a-f

určují počet započítávaných bodů, je-li heslo nalezeno v příslušných normalizovaných polích v SOLR (odpovídající pole MARC jsou uvedena v tabulce).

a) title
b) annotace
c) subjects
d) konference

2) **Pole, ve kterém je dané slovo nalezeno v PSH**. Parametry a-d určují počet bodů započítávaných v případě, že je v záznamu dokumentu nalezena podoba hesla z příslušného normalizovaného PSH pole.
   a) cspreflabel: pole obsahující české preferované znění hesla PSH
   b) enpreflabel: pole obsahující anglické preferované znění hesla PSH
   c) csaltlabel: pole obsahující české nepreferované znění hesla PSH
   d) enaltlabel: pole obsahující anglické nepreferované znění hesla PSH
3) **Příslušnost k oboru (*proximity*).** Tímto parametrem jsou posilovány skupiny hesel ze stejných větví hesláře. Parametr přidává body navíc skupině hesel ze stejné větve PSH, přičemž hodnota je násobkem stanovené váhy a „odklonu od mediánu výskytu v dané větvi“ (INCAD, 2013, s. 5).
4) **Plná shoda dotazu (*exact*).** Parametr násobící ostatní body v případě, že je heslo v záznamu nalezeno přesně ve formě uvedené v PSH. U jednoslovných hesel je započten vždy. Význam hraje tedy u víceslovných hesel, u kterých se v základním nastavení považuje za plnou shodu i případ, kdy je mezi slova hesla vloženo ještě jedno slovo. Tolerovaná vzdálenost se dá nastavit pomocí parametru „near“.
5) **Více slov dotazu (*multiple*).** Parametr pro víceslovná hesla, jejichž jednotlivé části jsou od sebe tak daleko, že nesplňují požadavky parametru plné shody (exact). Oproti plné shodě mívá mnohem nižší hodnotu a také se jím násobí zbývající body.
6) **Úroveň (hloubka) hesla ve stromové struktuře PSH (*level*)**. Tento parametr předpokládá, že čím hlouběji ve struktuře PSH je heslo nalezené v záznamu dokumentu, tím je přesnější a zaslouží si více bodů. Hodnota parametru je násobkem základní hodnoty parametru a úrovně hloubky, ve které se v PSH stromu heslo nachází.

Váhy jednotlivých parametrů se dají upravovat v uživatelském rozhraní nástroje. Lze například nastavit, aby se větší důraz kladl na hesla nalezená v názvu (title) a klíčových slovech (subjects) než na hesla v abstraktech a anotacích (annotation).

Nalezená hesla jsou seřazena podle počtu bodů označujících jejich relevanci k záznamu a dvě hesla s nejvyšším počtem bodů jsou přiřazena. Je-li možné přiřadit pouze jedno heslo, je přiřazeno samostatně.

Pokud nebyla nalezena žádná shoda PSH hesel s údaji v záznamu, je nutné přistoupit k přidělení „institučního hesla“ tj. hesla přidělovaného na základě názvu konference nebo oboru, kterému se věnuje instituce, která do NUŠL záznam dodala.

Výsledky projektu byly prezentovány zástupcem firmy Incad na 6. ročníku Semináře ke zpřístupňování šedé literatury v říjnu 2013. K tomuto datu bylo v repozitáři NUŠL Invenio 107 469 záznamů, z nichž 82 882 bylo neindexovaných, a byla na ně použita automatická indexace. Z počtu záznamů indexovaných automaticky bylo maximálně 1 455 záznamům přiděleno instituční heslo a 183 z nich nebylo možné zpracovat (Kocourek, 2013). V nezpracovatelných záznamech nebyla nalezena shoda s hesly PSH, neobsahovaly (použitelný) název konference a ani je nešlo zařadit podle zdrojové instituce, protože ta neměla jasně stanovený obor zájmu (to je obecně problém záznamů z vysokých škol, které mají široký záběr oborů).

Ve veřejné zakázce byl stanoven požadavek, že úspěšnost přidělování hesel musí být více než 60 %, přičemž úspěšností se rozumí, „že každý záznam bude obsahovat alespoň jedno odpovídající heslo PSH jakékoliv úrovně, naopak záznamy nebudou obsahovat neodpovídající hesla nebo hesla z úplně jiného stromu“ (Národní technická knihovna, 2012). Při počtech uváděných v říjnu 2013 byl tedy požadavek splněn, jelikož pouze k 1,75 % záznamů bylo přiřazeno instituční heslo a k 0,2 % žádné. Bohužel ale nebyla zveřejněna žádná analýza, která by ukazovala přesnost přiřazených hesel PSH.

Nástroje je i nadále využíván v současném provozu Národního úložiště šedé literatury tak, že je jednou za daný časový úsek spuštěn a záznamům jsou automaticky přidělena hesla PSH.

5.1.2.1. Problémy s automaticky přiřazenými hesly

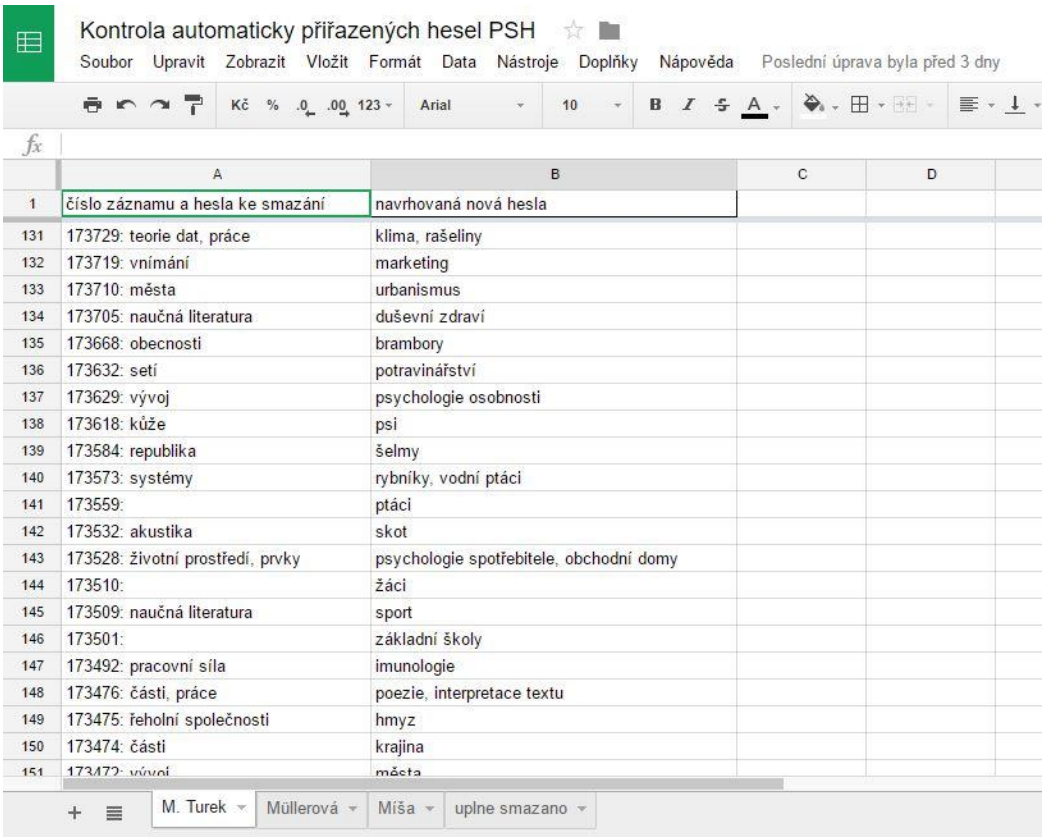
Automaticky přiřazená hesla jsou velkým přínosem pro celý systém NUŠL a i nadále jsou přiřazována novým záznamům. Přesto obsahují mnoho nepřesností a v současné době jsou tedy pro běžného uživatele v záznamu skryta, byť je možné podle nich vyhledávat a dále se s nimi pracuje při předávání dat do systému OpenGrey. Je možné, že výraznější nepřesnosti se začaly objevovat až po ukončení projektu vývoje nástroje, kdy byly váhy jednotlivých parametrů nastaveny tak, aby nástroj přiděloval co nejpřesnější hesla záznamům, které již v Inveniu byly,

ale toto nastavení již není tak dobře aplikovatelné na další záznamy, které od té doby přibyly. Je otázkou, do jaké míry lze tento problém řešit přenastavením vah, což by sice mohlo pomoci novým záznamům, ale určitě by uškodilo původním, respektive bylo by nutné nastavovat konkrétní váhy pro menší úseky než pro kompletní systém. Takové zásahy by ovšem vyžadovaly důkladnou analýzu, přepracování nástroje, a to vše s nejistým výsledkem.

V září roku 2014 byl v oddělení spravujícím NUŠL proveden pokus o podchycení a odstranění těch nejvýraznějších chyb, kterých se automatická indexace dopouštěla. Stážisti pracující v tomto oddělení Národní technické knihovny dostali za úkol kontrolovat automaticky přiřazená hesla. Tato práce mohla být svěřena stážistům, protože hlavním cílem nebyla podrobná kontrola a úprava všech jednotlivých automaticky naindexovaných záznamů, ale spíše vypátrání problematických hesel a oblastí. Konkrétně měli stážisti za úkol přejít na dodanou stránku s výsledky v systému Invenio a procházet záznamy. Pro tento účel byl připraven výstupní formát vyhledávání „html brief + PSH“, v němž zpracovatelé hned ve výpisu záznamů viděli tyto údaje:

- Název
- Autor
- Instituce
- Abstrakt
- Klíčová slova (přidělená ve zdrojové instituci)
- Klíčová slova (přidělená automatickou indexací)
- Číslo záznamu

Pokud zpracovatel na základě ostatních údajů v záznamu usoudil, že automaticky přiřazené heslo je chybné, zapisoval takové heslo spolu s číslem záznamu do tabulky. Důraz byl kladen na vyhledání zásadně problematických hesel, kdy je zřejmé, že došlo k chybě. Označená hesla byla určena ke smazání ze záznamu a k analýze ostatních záznamů s tímto automaticky přiřazeným heslem. Přestože bylo primárním cílem hledat chybná hesla, byli po seznámení s PSH někteří stážisti zároveň schopni k záznamům, u nichž určili heslo ke smazání, navrhnout i heslo PSH náhradní (ukázka tabulky na obr. 15).



Kontrola automaticky přiřazených hesel PSH

Soubor Upravit Zobrazit Vložit Formát Data Nástroje Doplňky Nápověda Poslední úprava byla před 3 dny

	A	B	C	D
1	číslo záznamu a hesla ke smazání	navrhovaná nová hesla		
131	173729: teorie dat, práce	klima, rašeliny		
132	173719: vnímání	marketing		
133	173710: města	urbanismus		
134	173705: naučná literatura	duševní zdraví		
135	173668: obecnosti	brambory		
136	173632: setí	potravinářství		
137	173629: vývoj	psychologie osobnosti		
138	173618: kůže	psi		
139	173584: republika	šelmy		
140	173573: systémy	rybníky, vodní ptáci		
141	173559:	ptáci		
142	173532: akustika	skot		
143	173528: životní prostředí, prvky	psychologie spotřebitele, obchodní domy		
144	173510:	žáci		
145	173509: naučná literatura	sport		
146	173501:	základní školy		
147	173492: pracovní síla	imunologie		
148	173476: části, práce	poezie, interpretace textu		
149	173475: řeholní společnosti	hmyz		
150	173474: části	krajina		
151	173472: vývoj	města		

M. Turek | Müllerová | Míša | uplne smazano

Obrázek 15: Tabulka pro kontrolu automaticky přiřazených hesel PSH

Celý proces neměl žádné akademické ambice, jednalo se skutečně spíše o drobný pokus o nápravu konkrétních problémů s automatickou indexací. Výsledky tak nejsou příliš reprezentativní a staly se spíše počátkem pro plánování dalších prací.

Na této kontrole pracovali postupně dva stážisti s různou odborností (stážista z ÚISK a dobrovolník po rekvalifikačním kurzu), kteří pracovali různě dlouho. Údaje o době zpracovávání nejsou k dispozici, ale celkově bylo zkontrolováno přibližně 2 000 záznamů a zhruba u 600 z nich došlo k nějaké korekci.

Pravděpodobně nejdůležitějším výsledkem bylo však vytipování hesel PSH, která bývají přiřazována chybně, a byla proto zařazena do slovníku stop slov. Jednalo se například o takováto hesla:

- „REPUBLIKA“ (Heslo se často se vyskytovalo v názvu, ale skutečně málokdy odpovídalo svému zařazení v PSH politologie > státověda > státní zřízení > republika.)

- „VÝVOJ“ (Problematické stejně jako většina dalších hesel v kategorii „obecnosti“, pod jedno heslo pak byl automaticky řazen jak ekonomický vývoj, tak vývoj plodu apod.)
- „VYUČOVACÍ PŘEDMĚTY“ (Heslo přiřazováno často kvůli řadě nepreferovaných termínů, které ovšem nejsou ekvivalenty, ale spíše druhovými termíny. Nedeskriptory jsou navíc poměrně obecné a z mnoha oblastí: kreslení, hudební výchova, psaní, cizí jazyk a dokument záznamu se tak málokdy věnoval pedagogice a vyučování.)

Jak již bylo zmíněno, nejedná se o reprezentativní data, na jejichž základě by došlo nebo mělo dojít k nějakým zásadním změnám v automatické indexaci v NUŠL. Hlavním výsledkem a zároveň důvodem, proč je pokus zmiňován v této práci, bylo rozhodnutí vyzkoušet, zda by v některých případech nebylo možné použít jiné a třeba i lepší řešení – jako například mapování klasifikačních schémat.

5.2. Využití sjednocení předmětového popisu NUŠL pro předání dat do OpenGrey

5.2.1. Představení systému OpenGrey

OpenGrey je mezinárodní evropská mezioborová databáze šedé literatury, systém v současné podobě a pod tímto názvem funguje od roku 2011. Provozovatelem je francouzský Institut vědeckých a technických informací INIST (Institut de l'Information Scientifique et Technique) sídlící ve Vandoeuvre-les-Nancy. V současné době do systému přispívá 16 institucí z 12 zemí[22], přičemž je ale nutné si uvědomit, že reálně se v systému objevují záznamy z mnohem většího množství institucí, jelikož partneři přispívající do databáze mohou působit i jako agregátoři (například NUŠL přispívá i záznamy z více než 120 institucí). Celkově nyní systém obsahuje celkem 1 014 821 záznamů v deseti různých jazycích (povinně je nutné uvádět anglický název a/nebo anglická klíčová slova) a všechny jsou označeny klasifikačním kódem SIGLE.

5.2.2. Historie systému

Systém OpenGray je nástupcem původního systému SIGLE provozovaného v letech 1980-2005 Evropskou asociací pro využívání šedé literatury EAGLE. Cílem SIGLE bylo shromažďovat a zpřístupňovat šedou literaturu ze zemí Evropského společenství (About OpenGrey, 2011). Provoz původního systému byl ukončen v roce 2005, přičemž už od února

[22] Partnerská instituce GreyNet je označena jako mezinárodní a tedy není zahrnuta do počtu zemí.

2005 nebylo možné přidávat nové záznamy a postupně byla databáze zakonzervována, protože už v té době se počítalo s možností jejího budoucího zveřejnění. Po definitivním ukončení činnosti EAGLE v roce 2006 se správcem databáze stal francouzský INIST, na jehož serveru byla databáze následně zveřejněna v roce 2007, tentokrát volně přístupná na internetu pod názvem OpenSIGLE, pod nímž fungovala na platformě DSpace až do roku 2011. V roce 2011 byl systém převeden z DSpace na software Exalead a znovu bylo umožněno přidávání záznamů i připojování dalších partnerských institucí. Jelikož došlo k velkým technologickým i obsahovým změnám a změnila se i celková politika databáze, bylo rozhodnuto o jejím přejmenování na dnešní OpenGrey. (Stock, 2011)

Historie tohoto systému je důležitá pro pochopení proměn a vývoje strategie věcného popisu v databázi, které se odvíjejí od naprosto odlišného ekonomického modelu, a tudíž i kooperace s partnerskými organizacemi dodávajícími záznamy v systémech SIGLE a OpenGrey.

SIGLE byl distribuován jako klasická placená databáze skrze databázová centra a na CD-ROM, zároveň zúčastněné partnerské instituce platily roční poplatek i poplatky za jednotlivé zveřejněné záznamy. Instituce produkující záznamy měly zájem o jejich zveřejnění v systému a dokonce za to platily, což bylo podstatné z hlediska uplatňování pravidel tvorby záznamů. Záznamy byly totiž vytvářeny či upravovány přímo pro tento systém podle jednotných pravidel vyžadujících použití jednotného klasifikačního schématu.

Naproti tomu mladší varianta systému OpenGrey pracuje na dobrovolné bázi a záznamy nejsou již vytvářené primárně pro tuto databázi a naopak se díky novým možnostem předávání záznamů (hlavně prostřednictvím protokolu OAI-PMH) jedná o záznamy až dodatečně upravované pro umístění do databáze OpenGrey. V takových případech už je pak nutné řešit sjednocování věcného popisu.

5.2.3. Klasifikační schéma

OpenGrey převzal ze systému SIGLE i klasifikační schéma SIGLE (SIGLE classification scheme), které bylo odvozeno z klasifikačního schématu americké Komise pro vědecké a technické informace COSATI (Committee on Scientific and Technical Information) spadající pod Federální úřad pro vědu a technologii (Federal Council for Science and Technology) (Schöpfel, 2007). Jedná se o poměrně hrubé, pouze dvouúrovňové klasifikační schéma obsahující 22 hlavních kategorií značených čísly 01 až 22. V druhé úrovni se pak dále dělí na 252 podkategorií značených kombinací označení hlavní větve spolu s písmeny (např. 08A,

22B). Kromě toho je v každé kategorii ještě jedna podkategorie obecného rázu, která není značená písmenem, ale nulou (například 080).

5.2.4. Mapování PSH na klasifikační schéma SIGLE

Podle pravidel pro předávání záznamů musí všechny záznamy předávané do OpenGrey obsahovat klasifikační kód SIGLE. Tím se přesouvá povinnost řešit sjednocování na partnerské instituce, které bývají i ve funkci národních agregátorů. Ty pak musí řešit nejen sjednocování svých vlastních agregovaných záznamů, ale i následný převod na klasifikační schéma SIGLE.

Národní úložiště šedé literatury je zástupcem v OpenGrey za Českou republiku, a aby mohly být záznamy předávány dále, bylo ve spolupráci s OpenGrey vytvořeno mapování mezi klasifikačním schématem SIGLE a PSH. PSH byl vybrán, jelikož se používá v Národní technické knihovně i v jejím institucionálním repozitáři a zároveň je využíván i u záznamů vkládaných přímo ručně do NUŠL.

Hesla PSH byla mapována pouze na 22 hlavních hrubých kategorií. Toto rozhodnutí mohlo být provedeno jak kvůli náročnosti takového mapování, tak kvůli poměrně nepřesným heslům PSH, která byla přidělena části záznamů automatickou indexací. Zobecněné kategorie k přiřazeným heslům mají pak větší šanci zakrýt takto vzniklé nepřesnosti.

Převod z PSH na SIGLE kódy probíhá v NUŠL v systému Invenio. V Inveniu je v rámci PSH indexeru uložena databázová tabulka obsahující pouze dva sloupce: sloupec s PSH identifikátorem ve formě PSHčíslo (např. PSH6486) a sloupec s označením kategorie SIGLE ve formě číselný_kód - textový popis (např. 05B - Information science, librarianship). Samotné převádění pak provádí pomocí tabulky skript, který se vždy „zeptá" na konkrétní heslo PSH posláním dotazu na URL http://invenio.nusl.cz/indexer/getos/[HESLO] a systém si dotaz převede na PSH ID a najde v tabulce odpovídající SIGLE kategorii. Například:

Dotaz: http://invenio.nusl.cz/indexer/getos/knihovny
Odpověď: 05B - Information science, librarianship

Dotaz pracuje s textovou podobou hesla PSH a ne s jeho ID, protože původně byla v NUŠL záznamech hesla PSH pouze v této podobě a ID byla ke každému heslu PSH v každém záznamu dodána až v roce 2014. V současné době je namapováno 13 482 PSH hesel z celkového počtu 14 130 hesel, takže chybí mapování pro cca 500 hesel. Ani to však není pro mapování problém, jelikož systém pak automaticky najde nadřazený termín nenamapovaného hesla a dále pracuje s ním.

6. Využití mapování ke sjednocování věcného popisu v systému NUŠL

Po představení zahraničních řešení problému sjednocování věcného popisu v repozitářích (1. kapitola) a seznámení se systémem NUŠL (2. kapitola) se dostáváme k hlavnímu tématu této diplomové práce, jímž je možnost využití mapování jako metody sjednocování věcného popisu v systému NUŠL.

V systému NUŠL již byly prováděny pokusy s automatickou indexací za pomoci strojového učení a i v současnosti je aplikována automatická indexace založená na vyhledávání a porovnávání. Používaná automatická indexace sice hodně využívá věcných údajů, které byly do záznamu přiřazené ve zdrojové instituci, nicméně tato práce si klade otázku, zda by v případech, kdy jsou dokumenty již věcně zpracovány, nebylo vhodnější použít ke sjednocení věcného popisu mapování. Stanovená hypotéza předpokládá, že výsledný věcný popis záznamů bude po použití mapování přesnější než s využitím automatické indexace. Výchozím předpokladem je, že mapování je lepším využitím již jednou vykonané intelektuální činnosti odborníků než automatická indexace. Velká výhoda zároveň spočívá v tom, že zatímco záznamy indexované automaticky by měly být optimálně vždy ještě zkontrolovány člověkem (což vzhledem k množství záznamů v digitálním repozitáři NUŠL není možné), u mapování je lidský zásah vlastně předsunut před samotnou indexaci cílovým selekčním jazykem a je nutný v zásadě pouze jednou. Jako cílový selekční jazyk pro mapování byl zvolen PSH, který již byl k účelu sjednocování věcného popisu použit při automatické indexaci.

V následující části bude představen základní princip mapování, které jsem provedla v rámci této práce. Následně budou vždy po krátkém představení mapovaného schématu a jeho struktury popsány konkrétní postupy při mapování na PSH a výsledky aplikace mapování na danou množinu záznamů.

V rámci této práce bylo na PSH namapováno schéma Konspekt a tezaurus MeSH, jejichž údaje se vyskytují v záznamech Národní lékařské knihovny (NLK), které jsou přebírány do systému NUŠL. Na tyto záznamy bylo následně mapování aplikováno a testováno. Kromě toho byly analyzovány možnosti využití mapování u záznamů VŠKP na případu tří vysokých škol předávajících záznamy do systému NUŠL.

6.1. Mapování

Mapování je v rámci této práci chápáno ve významu definovaném normou ISO 25964 jako „proces vytváření vztahů mezi pojmy jednoho a druhého slovníku“ (ISO 25964-2, 2013, s. 7) a jako produkt takového procesu.

Na základě mapování vytvořených v rámci této práce budou do záznamů sklízených do repozitáře NUŠL Invenio přidělena hesla PSH a zapsána do příslušného pole. Pro odlišení od hesel PSH přidaných ručně nebo automatickou indexací budou mít pole obsahující hesla přidaná mapováním jinou hodnotu druhého indikátoru.

Mezinárodní norma ISO 25964 vyšla ve dvou částech v letech 2011 a 2013. Zatímco první část celkově pokrývá vývoj a správu tezaurů, druhá část se věnuje vzájemnému mapování tezaurů (a dalších „slovníků“ – nástrojů věcného popisu) pro zajištění interoperability. Jedná se o výchozí metodický materiál, přestože se cíle normy i mapování v systému NUŠL částečně liší. Norma totiž pracuje hlavně s předpokladem, že vytvořená mapování budou využita v nějaké formě federativního vyhledávání, zatímco v případě NUŠL bude mapování použito k obohacení záznamů, a tím se přispěje ke snaze o sjednocení věcného popisu záznamů v repozitáři. Základní informace týkající se mapování samotného nicméně zůstávají stejné v obou případech.

Mapování je vytvářeno mezi pojmy a termíny různých slovníků. Slovníkem jsou v případě této normy myšleny tezaury, klasifikační schémata, taxonomie, jmenné autority a ontologie, o nichž se zmiňuje, nicméně nestanovuje konkrétní definici termínu. Termín pojem[23] je tu definován jako „jednotka myšlenky“ (ISO 25964-2, 2013, s. 4) a funguje jako synonymum k termínům třída nebo kategorie. Samotné slovo termín vyjadřuje „slovo nebo slovní spojení používané k označení pojmu“ (ISO 25964-2, 2013, s. 14).

Do procesu mapování vstupují dva druhy slovníků. Zdrojový slovník, který norma definuje jako „slovník sloužící jako počáteční bod při hledání odpovídajícího termínu nebo pojmu jiného slovníku,“ a cílový slovník, „v němž je hledán termín nebo pojem odpovídající existujícímu termínu nebo pojmu ve zdrojovém slovníku“ (ISO 25964-2, 2013, s. 13). Cílovým slovníkem je v tomto případě PSH.

[23] Norma používá anglický termín „concept“, který je dle české TDKIV překládán jako „pojem“.

Mapování jako produkt se bude skládat z termínu nebo jiného označení pojmu zdrojového slovníku, jednoho až dvou trvalých identifikátorů URI odpovídajících hesel cílového slovníku (PSH) a kódu označení vztahu mezi mapovanými termíny. Použití identifikátoru URI zajistí použitelnost mapování i při změně používaného preferovaného termínu v PSH a usnadní další strojové zpracování.

Vyznačování vztahů není povinný požadavek normy, je to však dobrý zdroj informací o kvalitě mapování pro závěrečné vyhodnocení. Označení jsou uváděna ve smyslu označení vztahu hesla PSH k heslu zdrojového slovníku. Zapisované vztahy označují jednak míru ekvivalence, jednak hierarchie. Vychází částečně ze zmiňované normy, částečně ze specifikace SKOS (World Wide Web Consortium, 2009). Vztahy jsou seřazené podle priority, s níž byly vyhledávány pro mapování.

a) Přesná shoda (e – exact match). Tento vztah značí vysokou úroveň shody, kdy jsou pojmy z mapovaných slovníků s velkou jistotou zaměnitelné.
b) Blízká shoda (c – close match). Vychází ze specifikace SKOS a je používán, pokud je za určitých okolností možné použít mapované pojmy zaměnitelně, ale zároveň se nedá tato shoda uznat za přesnou.
c) Nadřazenost (b – broad match). Označení určuje, že mapovaný pojem PSH je hierarchicky nadřazený mapovanému pojmu ze zdrojového slovníku. Heslo PSH je pak ve znění normy charakterizováno jako „nadřazený termín“, tedy „preferovaný termín reprezentující pojem, který je širší než pojem dotazovaný“ (ISO 25964, 2013, s. 4). TDKIV definuje nadřazený termín srozumitelněji jako „termín označující rozsahově širší pojem, který v hierarchickém vztahu reprezentuje třídu nebo celek“ (Balíková, 2003e).
d) Podřazenost (n – narrow match). Funguje obdobně jako vztah nadřazenosti s tím rozdílem, že tentokrát je pojem PSH hierarchicky podřízen druhému pojmu. Rozsah podřazeného pojmu musí dle normy plně spadat do rozsahu nadřazeného termínu. Podřazený termín je v TDKIV definován jako „lexikální jednotka označující rozsahově užší pojem, který v hierarchickém vztahu reprezentuje členy třídy nebo části celku“ (Balíková, 2003f).

Většina vytvořených mapování bude typu „jeden-na-jeden[24]“ (one-to-one), ve kterém „jediný pojem jednoho slovníku je mapován na jediný pojem slovníku druhého“. Zároveň ovšem není vyloučeno (a norma na tuto možnost upozorňuje), že jeden pojem může mít vícero mapování tohoto typu, pokud na sobě nejsou nijak závislá.

Menší část mapování bude typu „jeden-na-mnoho“ (one-to-many), což je „mapování pojmu z jednoho slovníku na kombinaci dvou a více pojmů jiného slovníku“. V našem případě je stanoven maximální počet pojmů na dva, takže na pojem jiného slovníku může být namapována kombinace nanejvýš dvou hesel PSH.

[24] V literatuře se překládá i jako jeden-na–jednoho, nicméně v této práci budu používat jeden-na-jeden, jelikož je mapován jeden pojem na jeden pojem. Nejednoznačný je i způsob zápisu, zde byla zvolena varianta se spojovníky.

6.2. Konspekt

Konspekt[25] je metoda pro popis a hodnocení fondů knihoven vytvořená v 70. letech Skupinou výzkumných knihoven (Research Libraries Group, zkráceně RLG). Tato metoda byla vyvinuta v reakci na informační explozi druhé poloviny 20. století, která změnila akviziční politiku knihoven (Balíková, 2003g). Fondy a akvizici již nebylo možné a žádoucí hodnotit podle množství informačních zdrojů, ale spíše podle plnění potřeb uživatelů knihovny. Zároveň nabyla na významu kooperativní akvizice, a bylo tedy zapotřebí nástroje umožňujícího předávání informací o složení fondu mezi knihovnami.

Konspekt představuje řadu evaluačních technik a softwarových nástrojů pro popis a hodnocení fondu. Jednou z hlavních komponent je i kategorizační schéma, do jehož kategorií jsou podle obsahu řazeny jednotlivé informační zdroje knihovních fondů. S pomocí těchto součástí umožňuje Konspekt charakterizovat fond z pěti hledisek (Presová, 2005 s. 16):

1. Současný stav fondu: současné pokrytí jednotlivých tematických oblastí ve fondu.
2. Současná akviziční politika: úroveň současného doplňování fondu jednotlivých tematických oblastí.
3. Strategie budování fondu.
4. Jazykové pokrytí: informace o jazykovém složení fondu.
5. Ochrana fondu: plánované činnosti ochrany fondu.

Všechna tato hlediska se po vyhodnocení popisují pomocí indikátorů, což zjednodušuje mezinárodní srozumitelnost a umožňuje srovnání mezi knihovnami.

6.2.1. Konspekt v ČR

V České republice se o metodu Konspektu a její lokalizaci do českého jazyka a českého prostředí stará Národní knihovna ČR. Ta už v září 2001 začala testovat přidělování skupin Konspektu do bibliografických záznamů a od ledna 2002 se již jednalo o běžnou součást nových záznamů (Presová, 2005 s. 54). V současné době je skupina Konspektu povinným selekčním prvkem již v minimálních záznamech NK ČR i pro záznamy předávané do Souborného katalogu ČR (Balíková, 2012). Zároveň se konspektového schématu využívá i v Jednotné informační bráně pro tematické třídění informačních zdrojů.

[25] Konspekt je v češtině zapisován ve variantě s malým i velkým počátečním písmenem, v této práci budu používat verzi „Konspekt“.

Pro zápis údajů Konspektu do bibliografických záznamů platí daná pravidla o maximálním počtu přiřazených skupin Konspektu. Do každého záznamu je s ohledem na primární funkci schématu (tj. k využití k hodnocení fondů) možné vložit pouze jednu a ve výjimečných definovaných případech dvě skupiny Konspektu. Dvě skupiny je možné použít pouze v případě, kdy jedna se týká obsahu a druhá formy dokumentu záznamu (Literatura pro děti a mládež, učebnice, jazykové slovníky, bibliografie, biografie, rukopisy, staré tisky a vzácné dokumenty) (Lichtenbergová, 2000). Obsahová skupina by měla být zapisována jako první v pořadí.

Do záznamů samotných jsou údaje o odpovídajících skupinách Konspektu zapisovány do pole 615 v případě formátu UNIMARC, nebo do pole 072 v případě MARC 21. Samotné pole pak obsahuje povinně podpole $n obsahující notaci MDT a podpole $a se slovním vyjádřením skupiny Konspektu.

6.2.2. Schéma Konspektu

Pro tuto práci je důležité zmiňované kategorizační schéma Konspekt, které je v českém prostředí používáno i jako selekční jazyk jak v klasických knihovních katalozích, tak i v systémech zaměřených na zpřístupňování elektronických informačních zdrojů. Je ovšem důležité mít na paměti, že jde původně o kategorizační schéma vytvořené pro popis a hodnocení fondu.

Původní schéma Konspektu, dnes označované jako RLG Konspekt[26], vzniklo ve spolupráci RLG a Asociace vědeckých knihoven (ARL), takže odpovídalo zejména potřebám velkých amerických vědeckých knihoven. Toto dvouúrovňové schéma má 25 hlavních „předmětových kategorií“ a přes 8 000 „předmětů“, které jsou navázány na klasifikační znaky Třídění kongresové knihovny (LCC) (Presová, 2005 s. 17).

Pro další typy knihoven, do nichž se rozšířila metoda Konspektu, se ovšem toto schéma ukázalo jako nevyhovující a vznikla upravená verze tzv. WLN Konspekt. Tato varianta již byla tříúrovňová:

1. 24 předmětových kategorií
2. 500 skupin Konspektu
3. 4 000 předmětů – vázané na notaci DDT a slovní vyjádření dané notace

[26] Označení RLG Konspekt a WLN Konspekt převzato z publikace PRESOVÁ, Silvie, 2005. *Metoda Konspektu a její vztah k selekčním jazykům: současný stav a trendy se zaměřením na situaci v České republice.*

Při rozšiřování metody Konspektu do dalších typů knihoven a zvláště do dalších zemí dochází k úpravám schématu pro potřeby dané knihovny. Nejinak tomu bylo i při zavádění Konspektu do českého prostředí.

Česká varianta schéma Konspektu vychází z WLN Konspektu v tom, že má úroveň „skupin Konspektu“. Jinak se ale jedná pouze o dvouúrovňovou hierarchii:

1. 26 předmětových kategorií: slovně vyjádřené a označené pořadovým číslem
2. 605 skupin Konspektu: skládá se z notace MDT a odpovídajícího slovního vyjádření

Třetí hierarchická úroveň jako taková u nás nebyla vytvořena, nicméně do záznamů tematických autorit v Databázi národních autorit NK ČR jsou taktéž přidávány údaje o skupině Konspektu, do níž tematicky patří. Tato data lze teoreticky využít jako třetí úroveň – předmětů Konspekt, nicméně momentálně tak využívána nejsou.

Další rozdíly mezi původními schématy a tím českým se odvíjí od toho, že zatímco americká schémata Konspekt byla vázána na DDT a LCC, v české verzi se přistoupilo k přepracování do MDT, které se používá u nás. Byly tak zrušeny skupiny Konspektu, jejichž notace DDT nemá paralelu v MDT a v souvislosti s MDT došlo i k dalším úpravám rozšíření slovních vyjádření (Balíková, 2003g). Zásadní pak byly změny nutné pro aplikaci schématu do českého prostředí, neboť v původních verzích Konspektu nebyla rozpracována některá pro nás zásadní témata (česká literatura, čeština).

Konspektové schéma bývá označováno jako kategorizační schéma, které je definováno jako „seznam kategorií (skupin, tříd), sloužící k seskupování organizovaných entit na základě jejich příslušnosti k určité kategorii“ (Bratková, 2014).

Jak již ovšem bylo zmíněno, v českém prostředí se používá schéma Konspektu nejen jako součást nástroje pro hodnocení fondu, ale i jako selekční jazyk v katalozích i webových službách pracujících s informačními zdroji.

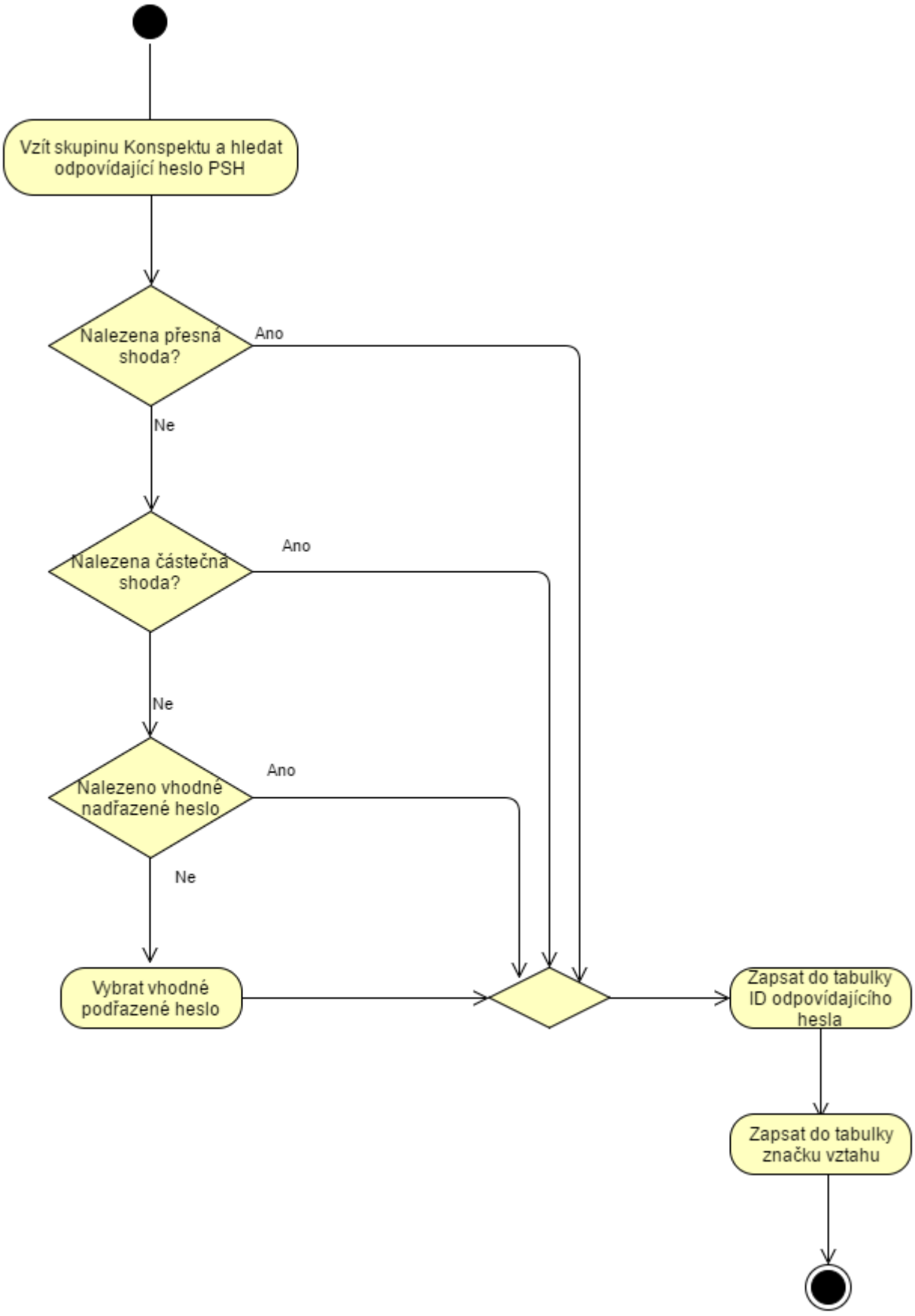
Malá hierarchická hloubka tohoto schématu má za následek, že oproti klasickým selekčním jazykům poskytuje možnost jen poměrně hrubého třídění a že jsou často na stejné úrovni skupiny, které by měly být vůči sobě nadřazené/podřazené. V rámci schémat Konspektu jsou řešeny pouze vztahy hierarchické. Neřešení vztahů ekvivalence se pak odráží v absenci řízeného slovníku a v důsledku toho i v omezené pojmové základně. Nejsou řešeny ani vztahy asociační a vzhledem k indexačním pravidlům pro práci s Konspektem, která omezují počet

přiřazovaných skupin Konspektu do záznamu, není možné takové vazby vyjádřit ani přímo v bibliografických záznamech.

6.2.3. Mapování Konspektu

Cílem bylo, aby každá skupina Konspektu měla přiřazené odpovídající heslo PSH nebo kombinaci dvou a zároveň aby byl určen druh vztahu mezi mapovanými hesly.

Data o aktuální podobě schématu Konspektu byla získána z webových stránek NK ČR, odkud bylo vzhledem k malé velikosti možné překopírovat schéma do tabulkového editoru. Po vyčištění dat jim byla postupně ručně přiřazována URI ID ČR hesel PSH, která byla vyhledávána v rozhraní prohlížení PSH. Vyhledáváno bylo nejprve heslo, jehož pojem přesně odpovídal dané skupině Konspektu. Pokud takové heslo nebylo k dispozici, pokračovalo se s hledáním částečné shody nebo nadřazeného hesla. Výjimečně byla přiřazena hesla podřazená. Postup mapování je naznačen na diagramu aktivit (typ UML diagramu) na obr. 16.



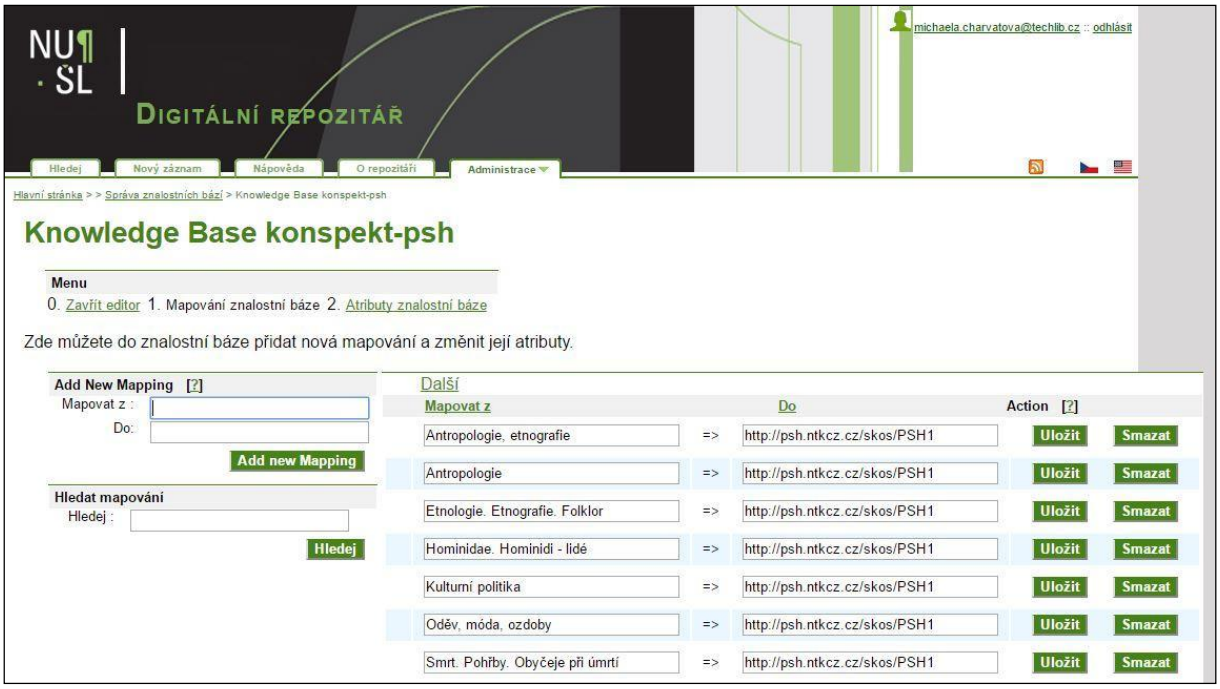
Obrázek 16: Diagram aktivit procesu mapování skupin Konspektu

Tímto procesem vznikla tabulka s kombinacemi údajů: skupina Konspekt, PSH ID odpovídajícího hesla, případně ID 2. hesla PSH a kód hodnoty shody. Podobu tabulky lze vidět na obr. 17.

	A	B	C	D	E
1			Shoda	PSH	PSH2
2	1 -	Antropologie, etnografie	e	http://psh.ntkcz.cz/skos/PSH1	
3		Antropologie	c	http://psh.ntkcz.cz/skos/PSH1	
4		Etnologie. Etnografie. Folklor	b	http://psh.ntkcz.cz/skos/PSH1	
5		Folklor	e	http://psh.ntkcz.cz/skos/PSH70	
6		Hominidae. Hominidi - lidé	b	http://psh.ntkcz.cz/skos/PSH1	
7		Kulturní politika	b	http://psh.ntkcz.cz/skos/PSH1	
8		Oděv, móda, ozdoby	b	http://psh.ntkcz.cz/skos/PSH1	
9		Smrt. Pohřby. Obyneje při úmrtí	b	http://psh.ntkcz.cz/skos/PSH1	
10		Sociologie kultury. Kulturní život	b	http://psh.ntkcz.cz/skos/PSH1	
11		Spolenenské chování. Etiketa	b	http://psh.ntkcz.cz/skos/PSH1	
12		Veřejný a spolenenský život. Každodenní život	b	http://psh.ntkcz.cz/skos/PSH1	
13		Zvyky, mravy, obyneje v soukromém životě	b	http://psh.ntkcz.cz/skos/PSH1	
14					
15	2 -	Biologické vědy	n	http://psh.ntkcz.cz/skos/PSH573	
16		Biochemie. Molekulární biologie. Biofyzika	b	http://psh.ntkcz.cz/skos/PSH573	
17		Biologické vědy	n	http://psh.ntkcz.cz/skos/PSH573	
18		Biotechnologie. Genetické inženýrství	n	http://psh.ntkcz.cz/skos/PSH806	
19		Botanika	e	http://psh.ntkcz.cz/skos/PSH854	
20		Buněnná biologie. Cytologie	e	http://psh.ntkcz.cz/skos/PSH630	
21		Mikrobiologie	e	http://psh.ntkcz.cz/skos/PSH826	

Obrázek 17: Pracovní tabulka mapování skupin Konspektu

Z takto vytvořené tabulky byla data následně nahrána do modulu BibKnowledge systému Invenio (viz obr 18.). Do tohoto modulu je možné nahrát řadu „znalostních bází“, které pak mohou být v systému různě využívány. Z této konkrétní báze budou data použita při konverzi záznamů přebíraných přes protokol OAI-PMH do NUŠL Invenio, kdy budou do záznamů místo skupin Konspektu přidávána namapovaná hesla PSH. Do bibliografického záznamu v NUŠL bude dodáno pole 650 s podpoli obsahujícími české a anglické slovní vyjádření, PSH ID a údaj, že se jedná o heslo z PSH. Všechny tyto úkony jsou po zapracování do konverzního skriptu prováděny automaticky. V pilotní verzi nebyla hesla PSH dodaná skrze mapování nijak speciálně označena, ale v budoucnu bude k rozlišení použit první indikátor pole, který se dnes již využívá k rozlišení ručně a automaticky přidaných hesel.



Obrázek 18: Modul BibKnowledge s otevřenou bází pro mapování Konspekt-PSH

6.2.3.1. Aktualizace

Při používání mapování jako nástroje pro sjednocování věcného popisu v repozitáři je nutné počítat s tím, že mapovaný zdrojový slovník se může měnit, a mít proto připraven postup pro řešení takové situace.

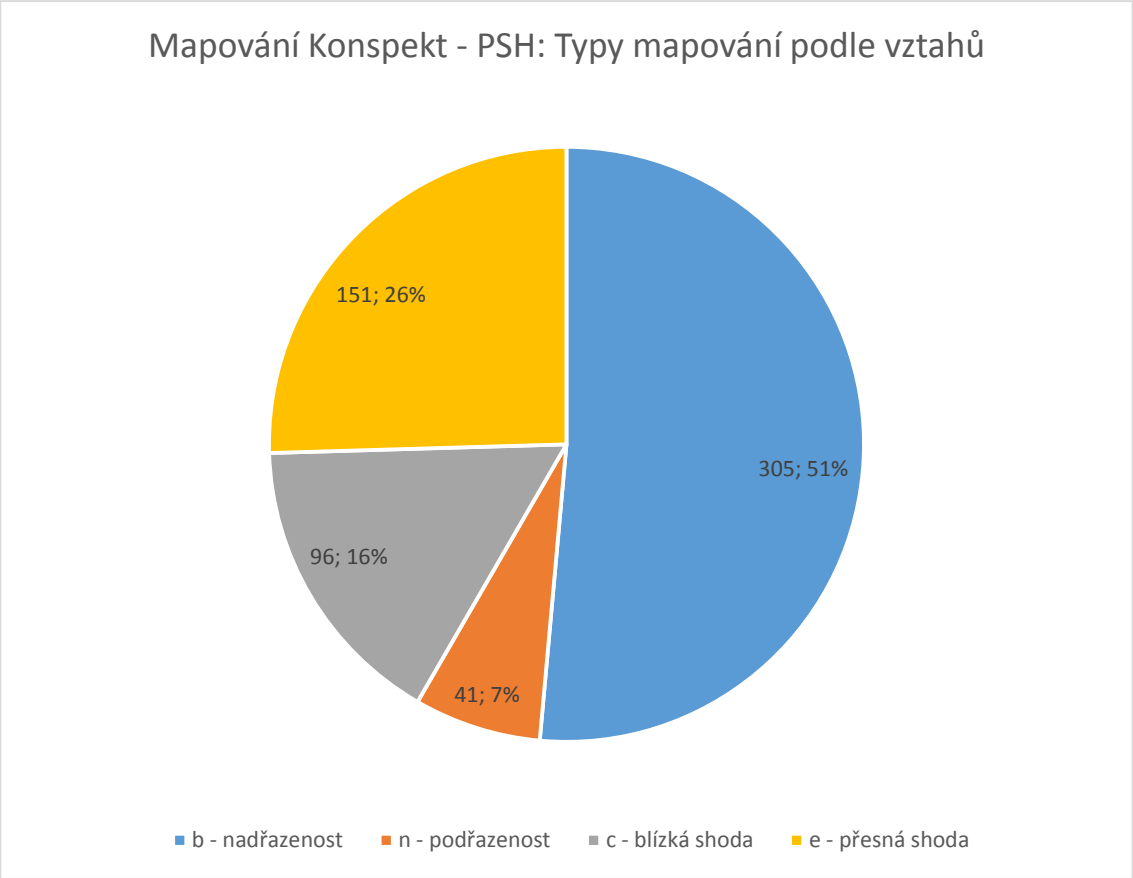
Informace o aktualizacích české verze předmětového schématu Konspekt jsou k dispozici na stránkách Národní knihovny ČR, která ho spravuje. V případě potřeby je možné změny zapracovat přímo v systému Invenio v modulu BibKnowledge, případně jednoduše pozměnit výchozí tabulku a tou následně kompletně přepsat existující znalostní bázi v Inveniu. Obdobně se bude postupovat i v případě zásadních změn v PSH.

Označení hesel přiřazených na základě mapování umožní případné změny provádět v záznamech i bez nutnosti opětovného sklizení a nové konverze záznamů. Stejně tak je pro možnost provádění automatizovaných změn v již sklizených záznamech důležité i podpole s identifikátorem PSH.

6.2.3.2. Charakteristiky mapování schéma Konspektu – PSH

V rámci procesu mapování byla vytvořena propojení mezi PSH a předmětovými kategoriemi i skupinami Konspektu. Kategorie Konspektu se nepoužívají v záznamech a jejich mapování bylo spíše pomůckou pro další práci; vzhledem k tomu je důležité hlavně mapování skupin Konspektu.

Bylo namapováno 593 skupin Konspektu z 26 kategorií Konspektu. Pro vyhodnocení procesu mapování je možné využít data o vztazích mezi jednotlivými mapovanými hesly, která zároveň vypovídají i o překryvu mapovaných schémat. Nejčastěji bylo vytvořeno mapování, kdy bylo přiřazené heslo PSH obsahově nadřazené mapované skupině Konspektu, a to v 305 případech (51,4 % celkového počtu mapování). Oproti tomu nejnižší počet (41 případů, 6,9 %) mapování tvoří skupina „n" vyjadřující vztah, ve kterém přiřazené heslo PSH pokrývá pouze část obsahu zahrnutého skupinou Konspektu. Vztah přesné a blízké shody mezi mapovanými hesly je dohromady více než 40 %. Tyto údaje přehledně zobrazuje graf na obr. 19.

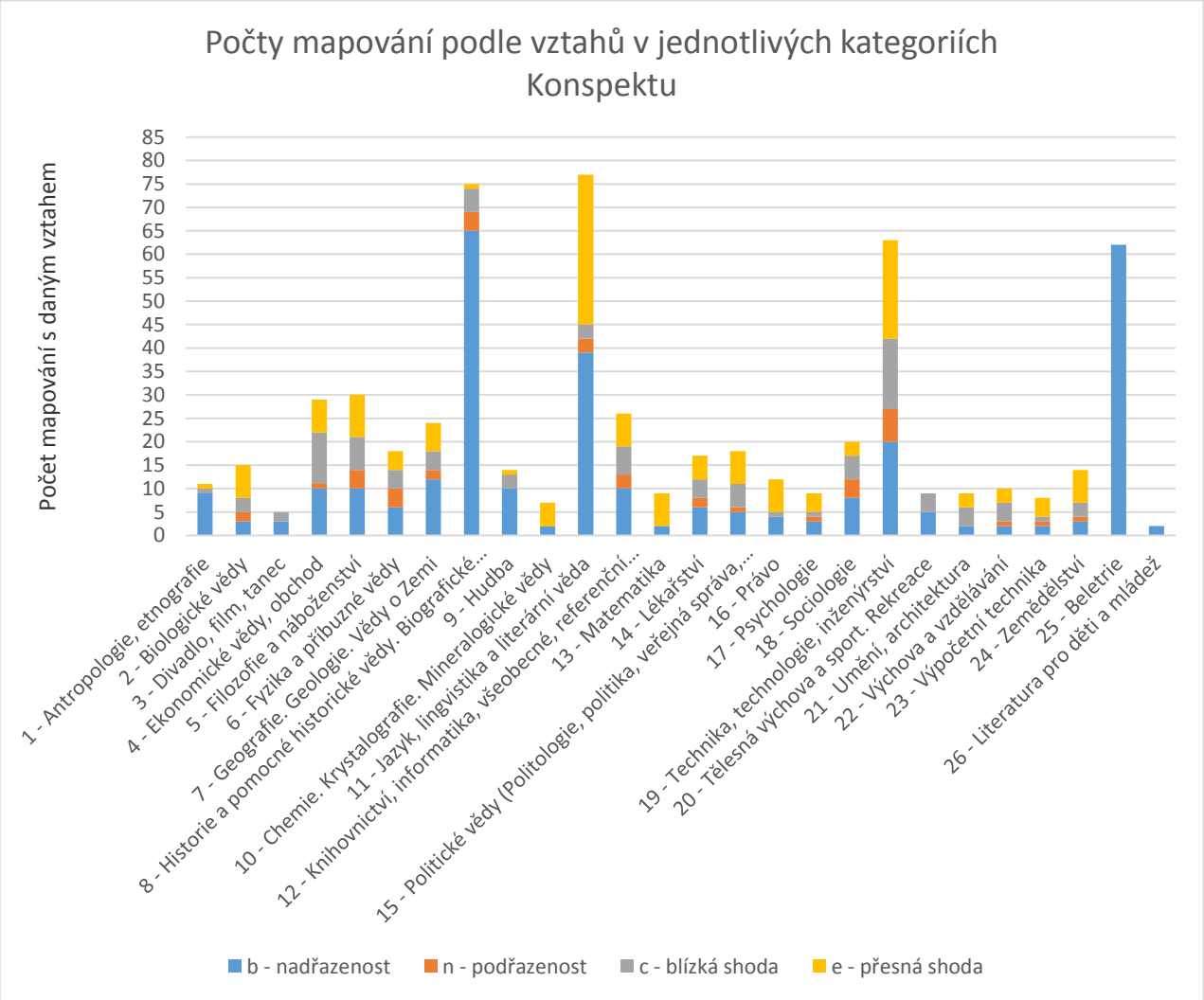


Obrázek 19: Typy mapování podle vztahů v mapování Konspekt - PSH

Zajímavé je nerovnovážné rozložení jednotlivých typů mapování v jednotlivých kategoriích Konspektu, které je vyjádřeno v tabulce (tabulka 2) a následně vyjádřeno i grafem na obr. 20 a které ukazuje významné rozdíly mezi detailností propracování jednotlivých kategorií Konspektu. Přesto se nedá říci, že by rozdíly mezi poměry množství užití vztahů mapování vycházely pouze z nestejného množství skupin v jednotlivých kategoriích Konspektu. Významným faktorem bude i (ne)propracovanost PSH v některých tematických větvích a jeho primární zaměření na technické obory.

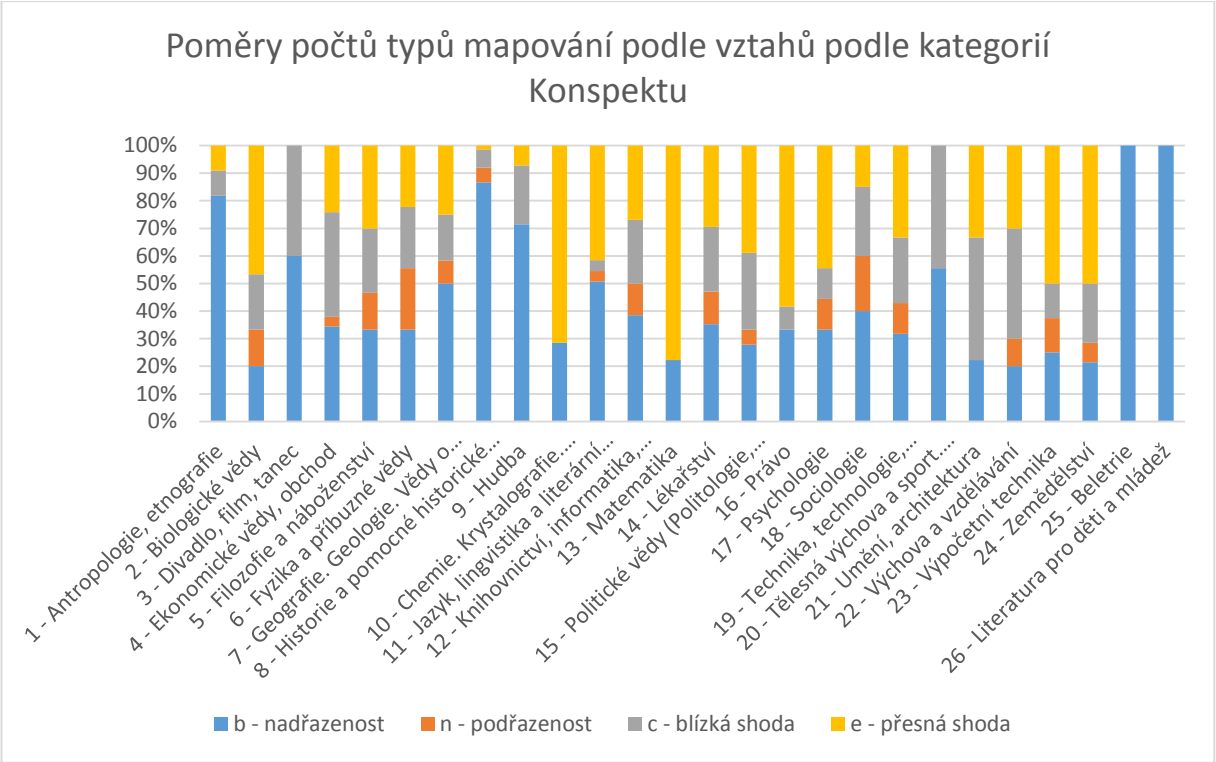
	1 - Antropologie, etnografie	2 - Biologické vědy	3 - Divadlo, film, tanec	4 - Ekonomické vědy, obchod	5 - Filozofie a náboženství	6 - Fyzika a příbuzné vědy	7 - Geografie. Geologie. Vědy o Zemi	8 - Historie a pomocné historické vědy. Biografické studie	9 - Hudba	10 - Chemie. Krystalografie. Mineralogické vědy	11 - Jazyk, lingvistika a literární věda	12 - Knihovnictví, informatika, všeobecné, referenční literatura	13 - Matematika	14 - Lékařství	15 - Politické vědy (Politologie, politika, veřejná správa, vojenství)	16 - Právo	17 - Psychologie	18 - Sociologie	19 - Technika, technologie, inženýrství	20 - Tělesná výchova a sport. Rekreace	21 - Umění, architektura	22 - Výchova a vzdělávání	23 - Výpočetní technika	24 - Zemědělství	25 - Beletrie	26 - Literatura pro děti a mládež
b - nadřazenost	9	3	3	10	10	6	12	65	10	2	39	10	2	6	5	4	3	8	20	5	2	2	2	3	62	2
n - podřazenost	0	2	0	1	4	4	2	4	0	0	3	3	0	2	1	0	1	4	7	0	0	1	1	1	0	0
c - blízká shoda	1	3	2	11	7	4	4	5	3	0	3	6	0	4	5	1	1	5	15	4	4	4	1	3	0	0
e - přesná shoda	1	7	0	7	9	4	6	1	1	5	32	7	7	5	7	7	4	3	21	0	3	3	4	7	0	0
celkem	11	15	5	29	30	18	24	75	14	7	77	26	9	17	18	12	9	20	63	9	9	10	8	14	62	2

Tabulka 2: Počty mapování podle vztahů v jednotlivých kategoriích Konspektu



Obrázek 20: Počty mapování podle typu vztahu v jednotlivých kategoriích Konspektu

Pro srovnání užití různých typů mapování podle vztahů bez ohledu na počet skupin Konspektu v jednotlivých kategoriích je určen graf na obr. 21.



Obrázek 21: Poměry počtů jednotlivých typů mapování podle vztahů v závislosti na kategorii skupiny Konspektu

Všechny skupiny kategorie „25 - Beletrie" jsou mapovány na heslo PSH „literatura" stojící vůči nim v nadřazeném vztahu, protože se jedná o skupiny určené pro formální specifikaci dokumentů národních literatur (česká literatura, anglická literatura atd.), nikoli o vyjádření jejich tématu. Obsah věnující se jednotlivým národním literaturám je označován skupinami Konspektu z kategorie „11 - Jazyk, lingvistika a literární věda", v níž jsou skupiny národních literatur označovány doplňkem „o ní" nebo „o nich", například „Česká literatura (o ní)" a „Baltské literatury (o nich)".

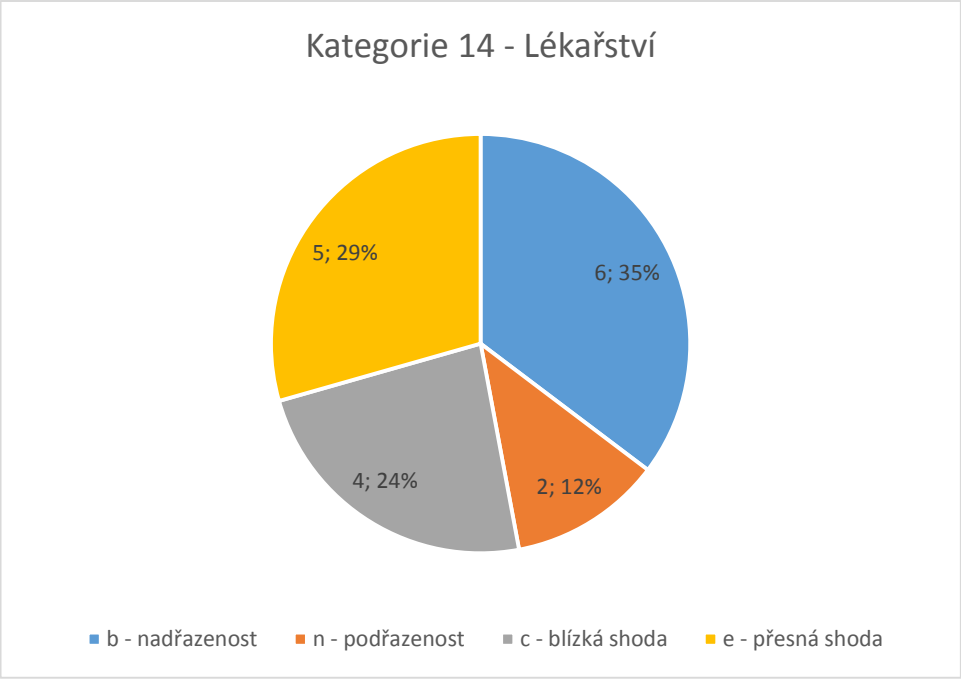
Obdobně bylo postupováno u kategorie „26 - Literatura pro děti a mládež", k jejímž skupinám bylo přiřazeno heslo PSH „Literatura pro děti a mládež". Přestože se jeví jako významově odpovídající, byl vztah označen jako „b", protože kategorie 26 je opět užívaná pouze k označování formy indexovaných dokumentů.

Kategorie „1 - Antropologie, etnografie", „8 - Historie a pomocné historické vědy. Biografické studie" a „9 - Hudba" se vyznačují nejvyšším poměrem počtu mapování typu „b" vůči celkovému počtu mapování v kategorii (více než 70 %). Vzhledem k tematickému zaměření kategorií lze předpokládat, že zde má velký vliv právě struktura těchto oblastí v PSH, která v humanitních vědách není příliš propracovaná a detailní.

Naproti tomu nejvyšší množství vztahů přesné shody mezi mapovanými hesly (také přesahující 70 %) najdeme v kategoriích „10 - Chemie. Krystalografie. Mineralogické vědy" a „13 - Matematika". To je odůvodnitelné nejen propracovaností těchto oborů v PSH (chemie je tematická větev PSH s největším počtem hesel), ale i poměrně jednoznačnou terminologií v chemii i matematice.

Vzhledem k tomu, že mapování Konspekt-PSH bylo pro testování využito ke zpracování záznamů NLK, je nejdůležitější kategorií „14 - Lékařství".

Kategorie „14 - Lékařství" obsahuje celkem 17 skupin Konspektu. Způsob namapování skupin je znázorněn v grafu na obr. 22, vyjadřujícím počet a procentuální poměr jednotlivých typů mapování podle vztahu.



Obrázek 22: Typy mapování v kategorii Konspektu 14 - lékařství

Vzhledem ke své důležitosti bude zde mapování této kategorie popsáno podrobněji. Tabulka č. 3 zobrazuje slovní vyjádření skupiny Konspektu, přiřazené heslo PSH včetně identifikátoru PSH a vztah mezi mapovanými pojmy.

Slovní vyjádření skupiny Konspektu	Vztah	PSH ID	Heslo PSH
Anatomie člověka a srovnávací anatomie	c	http://psh.ntkcz.cz/skos/PSH643	anatomie
Experimentální medicína	b	http://psh.ntkcz.cz/skos/PSH12931	lékařství
Farmacie. Farmakologie	c	http://psh.ntkcz.cz/skos/PSH13081	farmakologie
Fyziologie člověka a srovnávací fyziologie	c	http://psh.ntkcz.cz/skos/PSH12699	fyziologie člověka
Fyzioterapie. Psychoterapie. Alternativní lékařství	b	http://psh.ntkcz.cz/skos/PSH12748	terapie
Geriatrie	e	http://psh.ntkcz.cz/skos/PSH13028	geriatrie
Gynekologie. Porodnictví	n	http://psh.ntkcz.cz/skos/PSH13017	gynekologie
Hygiena. Lidské zdraví	n	http://psh.ntkcz.cz/skos/PSH13187	hygiena
Lékařské vědy. Lékařství	e	http://psh.ntkcz.cz/skos/PSH12931	lékařství
Ortopedie. Chirurgie. Oftalmologie	b	http://psh.ntkcz.cz/skos/PSH12931	lékařství
Patologie. Klinická medicína	b	http://psh.ntkcz.cz/skos/PSH12931	lékařství
Pediatrie	e	http://psh.ntkcz.cz/skos/PSH13023	pediatrie
Požáry. Ochrana před požáry	c	http://psh.ntkcz.cz/skos/PSH10636	požární ochrana
Psychiatrie	e	http://psh.ntkcz.cz/skos/PSH13050	psychiatrie
Stomatologie	e	http://psh.ntkcz.cz/skos/PSH12850	zubní lékařství
Úrazy a jejich prevence	b	http://psh.ntkcz.cz/skos/PSH13012	traumatologie
Veřejné zdraví a hygiena	b	http://psh.ntkcz.cz/skos/PSH13186	veřejné zdravotnictví

Tabulka 3: Tabulka mapování Kategorie Konspektu 14 - Lékařství

Problémy při mapování působilo hlavně sloučení několika širokých oborů a oblastí do jedné skupiny Konspektu. To platí nejen pro mapování skupin kategorie 14, ale i pro mapování všech ostatních kategorií. Nejvýrazněji se to projevuje na vyšším počtu mapování typu „b" - nadřazené heslo. V této kategorii nastává jen minimální počet případů, kdy by v PSH vůbec nebylo odpovídající heslo (výjimkou je např. skupina „experimentální medicína", pro kterou skutečně v PSH není ekvivalent). Často právě naopak máme v PSH několik samostatných hesel, která byla v Konspektu zařazena do jedné skupiny. Není možné přiřadit všechna dostupná hesla, jelikož při použití k indexaci by vždy 1-2 hesla byla do záznamu zařazena chybně. Zároveň není možné označit nadřazené heslo za blízkou nebo přesnou shodu, protože v PSH zahrnuje ještě další pojmy. Příkladem může být skupina „Ortopedie. Chirurgie. Oftalmologie", kde je zjevné, že případné přiřazení PSH všech tří hesel „ortopedie", „chirurgie" a „oftalmologie" ke každému záznamu by bylo chybné. Přiřazeno bylo tedy nadřazené heslo „lékařství", které zahrnuje i mnoho dalších lékařských oborů, a vztah byl tedy označen písmenem „b".

U skupin, které byly na hesla PSH mapovány ve vztahu „c", tedy blízká shoda, se většinou jednalo o případy, kdy byla skupina Konspektu rozšířena na dvě větší oblasti, které lze sice v PSH shrnout pod jedno heslo PSH (odpovídající jednomu z hesel označujících skupinu), ale existují důvody neoznačit takový vztah za přesnou shodu. Výjimkou je skupina „Požáry. Ochrana před požáry", která je překvapivě zařazena do kategorie lékařství. Nakonec bylo na skupinu namapováno heslo z nesouvisející větve PSH, které odkazuje na požární ochranu. V takovém případě se může zdát označení „c" - blízká shoda jako příliš silné vyjádření shody, nicméně po kontrole využívání této skupiny v katalogu NK ČR bylo zjištěno, že se v praxi využívá k indexaci všech dokumentů z oblasti hasičství apod. bez ohledu na zařazení do kategorie lékařství.

Přesná shoda byla využita pouze u lékařských oborů, které mají vlastní skupinu Konspektu (psychiatrie, stomatologie, geriatrie a pediatrie), a u skupiny obecné „Lékařské vědy. Lékařství".

Přestože byla snaha minimalizovat výskyt vztahu podřazenosti „n", byl v rámci této kategorie využit dvakrát. Vždy se k této variantě přistupuje, je-li lepší variantou než přiřazení nadřazeného hesla. V této kategorii byla podřazenost využita u skupiny „Gynekologie. Porodnictví", na něž bylo namapováno heslo PSH gynekologie, aby nemuselo být přiděleno obecné "lékařství". Na rozdíl od ostatních skupin Konspektu vyjadřujících více lékařských oborů se tyto obory překrývají a přidělení hesla vyjadřující pouze jeden z nich nebude chybné. Druhý výskyt vztahu „n" byl v případě skupiny "Hygiena. Lidské zdraví", kdy bylo přiřazeno heslo PSH „hygiena" zahrnující pouze část toho, co skupina Konspektu. Ani v tomto případě by využití vztahu podřazenosti nemělo působit problémy. Téma zdraví totiž není úplně ignorováno, jelikož ve stejné kategorii existuje ještě skupina „Veřejné zdraví a hygiena", mapována na nadřazený pojem „veřejné zdravotnictví".

Většina mapování je typu „jeden-na-jeden", protože mapování druhého typu „jeden-na-mnoho" bylo využito pouze ve třech případech:

- Geografie starověkého světa.
- Geologie. Meteorologie. Klimatologie.
- Historická věda. Pomocné vědy historické. Archivnictví.

6.2.3.3. Aplikace mapování na sadu záznamů

Mapování bylo aplikováno na 3 802 záznamů z Národní lékařské knihovny. Každý záznam obsahoval pouze jednu skupinu Konspektu, protože jich nebylo využito k určení formy popisovaného dokumentu. Celkový počet mapování se tedy shoduje s počtem záznamů a je jich také 3 802. Vzhledem k tomu, že mapování „jeden-na-mnoho“ bylo vytvořeno v tak málo případech a v oblastech nepokrytých v dokumentech, jejichž záznamy byly z NLK převzaty, není v testovací množině dat žádné takové mapování a počet hesel PSH přiřazených na základě mapování se rovná původnímu počtu skupin Konspektu v záznamech.

K původnímu popisu záznamů bylo celkově použito 26 různých skupin Konspektu, přičemž výrazně převažuje použití skupiny „Patologie. Klinická medicína“, která tvoří celých 45 % všech přiřazených skupin Konspektu. Dále je přiřazena tabulka 4 s přehledem nejčastěji použitých skupin Konspektu (vybrány jsou ty, které jsou užité alespoň 100krát) i s údajem o jejich procentuálním poměru užití v testované skupině záznamů.

Skupina Konspektu	Počet použití v množině záznamů	Procentuální podíl
Patologie. Klinická medicína	1 713	45,1 %
Lékařské vědy. Lékařství	649	17,1 %
Biochemie. Molekulární biologie. Biofyzika	325	8,5 %
Fyziologie člověka a srovnávací fyziologie	191	5,0 %
Ortopedie. Chirurgie. Oftalmologie	159	4,2 %
Pediatrie	118	3,1 %
Farmacie. Farmakologie	116	3,1 %

Tabulka 4: 7 nejčastěji užitých skupin Konspektu

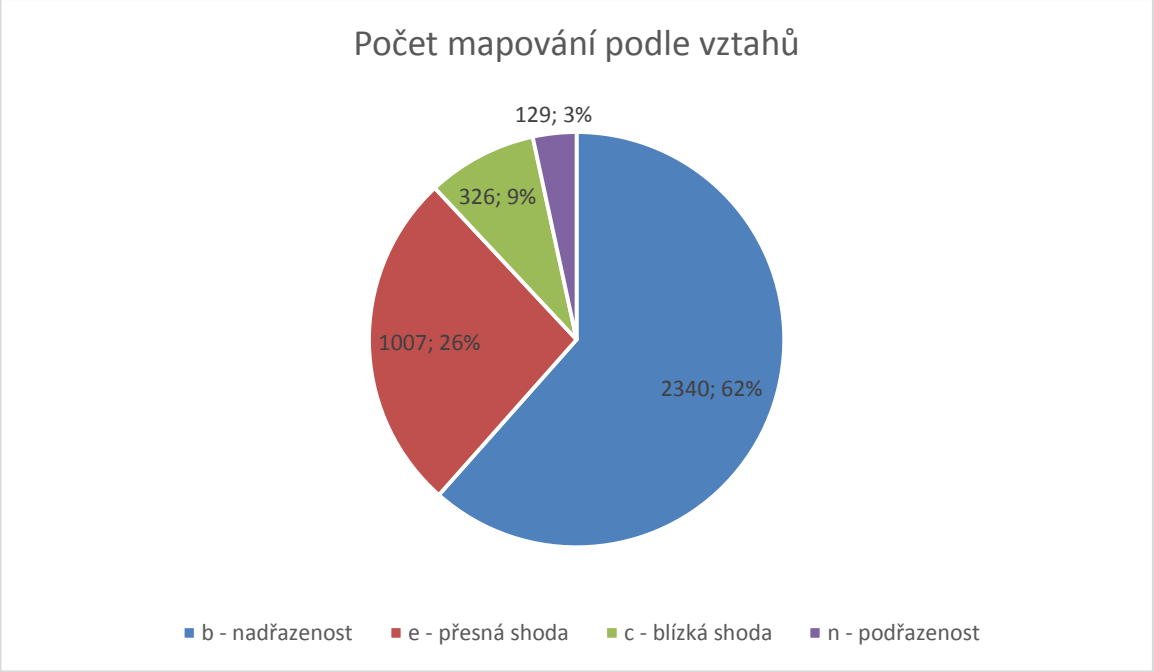
Na základě mapování skupin Konspektu bylo záznamům přiřazeno 33 různých hesel PSH. Nejčastější bylo podle očekávání heslo „lékařství“. Deset nejčastějších hesel PSH přiřazených na základě mapování skupin Konspektu je spolu s četností výskytu v tabulce 5:

Heslo PSH	Počet užití	Procentuální poměr
Lékařství	2 521	66,3 %
Biologie	337	8,9 %
fyziologie člověka	191	5,0 %
Pediatrie	118	3,1 %
farmakologie	116	3,1 %
zubní lékařství	94	2,5 %
veřejné zdravotnictví	81	2,1 %
gynekologie	78	2,1 %
psychiatrie	55	1,4 %
Genetika	38	1,0 %

Tabulka 5: 10 nejčastějších hesel PSH přidělených na základě mapování Konspektu

Z výše uvedených údajů vyplývá, že sice 66 % přiřazených hesel PSH tvoří obecné heslo „lékařství“, avšak na základě stejně obecného označení skupinou Konspektu „lékařství, lékařské vědy“ nebo z poměrně nepřesných hesel „Patologie. Klinická medicína“ a „Ortopedie. Chirurgie. Oftalmologie“.

Při pohledu na přiřazená hesla PSH z hlediska vztahu mezi mapovanými pojmy převládá vztah nadřazenosti a v tisíci případech vztah přesné shody. Heslo PSH bylo na základě blízké shody přiřazeno jen 326krát a heslo pokrývající rozsah jen částečně (vztah podřazenosti) jen 129krát. Poměry mezi počty vyjadřuje následující graf na obr. 25.



Obrázek 25: Graf počtu mapování podle typu vztahů po aplikaci na vybrané záznamy

6.3. MeSH

Medical Subject Headings (MeSH) je tezaurus zaměřený na oblast biomedicíny, který od roku 1960 vytváří americká Národní lékařská knihovna (Šímová, 2008). MeSH je dnes využíván v celé řadě odborných databází, knihovnách i lékařských centrech po celém světě.

Národní lékařská knihovna se věnuje překladu MeSH do češtiny již od roku 1977, kdy používala zkrácený překlad pro zpracování Oborové bibliografie BMČ (Šímová, 2008). Od roku 1991 pak vydává českou verzi v elektronické podobě a dále ji zpřístupňuje a distribuuje. Tezaurus je neustále aktualizován a pravidelně vychází nová verze jak anglického tezauru, tak jeho českého překladu. Překlad byl v minulosti soustředěn hlavně na preferované termíny (verze z roku 1991 obsahovala pouze překlad preferovaných termínů), ale v současné době jsou

překládány i nepreferované termíny, viz odkazy i definice (byť zde ještě není překlad oproti anglickému originálu kompletní). Tabulka 6 zobrazuje rozdíly mezi českou verzí a originálem ve verzi 2015 (Maixnerová, 2014):

	Český překlad	Anglický originál	Rozdíl
Preferované termíny	27 455	27 455	0
Nepreferované termíny	9 882	24 829	14 946
Viz odkazy	21 357	39 077	39 077
Definice	2 931	25 493	25 493

Tabulka 6: Rozdíly mezi českou a anglickou verzí MeSH 2015 (tabulka upravena podle Maixnerová, 2014)

Česká verze tezauru je k dispozici ve webovém prohlížení v portálu Medvik, na DVD nebo po podpisu licenční smlouvy ve strojově čitelné podobě ve formátu XML či ISO 2709.

6.3.1. Struktura tezauru

Verze 2015 obsahuje 27 455 deskriptorů[27], které jsou brány jako jednotka tezauru a odpovídají spíše tomu, co TDKIV označuje jako deskriptorový odstavec, tedy „hlavní části tezauru zahrnující deskriptor a všechny relevantní informace uvedené v poznámkovém a odkazovém aparátu, jako je definice, vysvětlující poznámka, poznámka o použití, nedeskriptory, podřazené, nadřazené a asociované deskriptory“ (Balíková, 2003h). V dokumentaci k českému překladu i v portálu Medvik jsou tyto jednotky označované jako „deskriptory“, budu se tedy v případě MeSH držet tohoto pojmenování.

Deskriptory jsou řazeny ve stromové struktuře 16 hlavních kategorií označených písmeny abecedy, které jsou základem notace jednotlivých deskriptorů. Kategorie se dále větví a s každým větvením je k notaci přidáno za tečku číselné označení:

[B] **Organismy**
- [B01] Eukaryota
  - [B01.050] zvířata
    - [B01.050.050] skupiny zvířecí populace...
    - [B01.050.150] Chordata...
    - [B01.050.500] bezobratlí
      - [B01.050.500.091] Annelida...

Každý deskriptor se může vyskytovat na více místech stromu a jeden deskriptor tak může mít přidělenu jednu a více notací podle počtu výskytů v tezauru.

[27] Uváděny jsou údaje k verzi z roku 2015, protože ta byla v rámci experimentu použita k mapování.

Kromě 16 tematických kategorií existuje ještě 17. kategorie Y pro podhesla – kvalifikátory, která slouží ke zpřesnění popisovaných témat při indexaci dokumentů. Podhesla mají vždy určené kategorie a podkategorie, k jejichž deskriptorům mohou být přiřazena.

6.3.2. Struktura záznamu deskriptoru MeSH

Záznam deskriptoru tezauru MeSH v portálu Medvik obsahuje tyto údaje (Národní lékařská knihovna, 2015) (označení jsou převzata přímo ze zobrazení záznamu):

1. Hlavní termín. Preferované termín v češtině.
2. Anglické záhlaví. Preferovaný termín v angličtině.
3. Viz. Nepreferované termíny v češtině.
4. Viz (anglicky). Nepreferované termíny v angličtině.
5. Související deskriptory. Odkazy typu viz též.
6. Preferovaný koncept. Odpovídá hlavnímu termínu.
7. Definice.
8. Další koncepty. Odkaz na záznam nedeskriptory, jejichž preferovaná i nepreferovaná znění jsou zařazena do Viz odkazů deskriptoru.
9. Anotace. Většinou se jedná o poznámky pro katalogizátory.
10. Poznámka k historii. Informace o vývoji deskriptoru v MeSH.
11. Online poznámka.
12. Veřejná poznámka.
13. Číslo záznamu. ID deskriptoru.
14. Povolená podhesla.
15. Zařazení v MeSH stromu (notace).

Obsah záznamu deskriptoru ve formátu odpovídá záznamu v portálu Medvik s tím, že některé údaje je třeba dále zpracovávat. Například zařazení ve stromu je rozděleno do jednotlivých částí a k získání kompletní notace je třeba tyto kousky pospojovat.

6.3.3. MeSH v bibliografickém záznamu

Bibliografické záznamy z NLK jsou do NUŠL přebírány pomocí protokolu OAI-PMH ve formátu MARCXML.

Údaje o přidělených deskriptorech MeSH přiřazených do bibliografického záznamu se nacházejí v opakovatelných polích 650, která mají v podpoli $2 hodnotu „czmesh“ (hesla

z jiných klasifikací mají v tomto poli svá vlastní označení). Příklad pole s MeSH deskriptorem v bibliografickém záznamu:

```
<datafield tag="650" ind1="0" ind2="7">
<subfield code="a">experimenty na zvířatech</subfield>
<subfield code="2">czmesh</subfield>
<subfield code="7">D032761</subfield>
</datafield>
```

Kromě podpole $2 označující použitý slovník obsahuje toto pole vždy podpole $a s textem preferovaného znění deskriptoru a podpole $7 s číslem záznamu deskriptoru.

6.3.4. Mapování MeSH – PSH

Před samotným vytvářením mapování mezi dvěma slovníky je nutné zjistit, jak bude možné aplikovat mapování do procesu konverze záznamů ze zdrojového repozitáře/katalogu. Pro rozhodování o podobě mapování je zásadní, jaké údaje o přiřazených heslech zdrojového slovníku k záznamům jsou dostupné v záznamech dodávaných přes protokol OAI-PMH. Dalšími rozhodujícími faktory jsou struktura mapovaných slovníků a forma, v jaké jsou tyto mapované slovníky dostupné.

Pro vytváření mapování v MeSH-PSH v rámci této práce bylo využíváno webové rozhraní Medvik, v němž je k dispozici český překlad MeSH ve stromové struktuře a český překlad MeSH ve strojově čitelném formátu MARCXML dostupný na vyžádání zdarma pro studijní a výzkumné účely. PSH byl využíván skrze rozhraní Prohlížení PSH a ve strojově čitelném formátu SKOS RDF/XML, ve kterém je dostupný volně ke stažení ze stránek NTK.

V bibliografických záznamech z NLK jsou údaje o přiřazených deskriptorech MeSH uvedeny v polích 650, přičemž záznam obsahuje vždy tolik polí 650, kolik má přidělených deskriptorů MeSH. V každém jednotlivém poli je uveden preferovaný termín, označení slovníku (czmesh) a číslo záznamu deskriptoru. Za takové situace je nejlepší vytvořit mapování mezi identifikátory hesel PSH a identifikátory záznamů deskriptorů, protože identifikátory jsou jednoznačné a nehrozí ani další problémy například s kódováním textů, které by se mohly objevit při mapování textových vyjádření hesel.

Číslo záznamu deskriptoru je základním identifikátorem v XML verzi tezauru, ale není nijak vázané na notaci ani jinak na umístění ve stromové struktuře. Ve webovém rozhraní je sice číslo záznamu uvedeno, ale nejedná se o identifikátor URL, ani o selekční údaj, podle něhož by

bylo možné záznam deskriptoru vyhledat. Ve webovém rozhraní je ale dobře vidět stromová struktura tezauru, která je užitečná pro proces mapování.

Mapování všech 27 000 hesel, pokud ho následně chceme aplikovat na necelé 4 000 bibliografických záznamů, by bylo skutečně neefektivní. Zároveň je jasné, že na deskriptory MeSH bude namapováno mnohem menší množství jedinečných hesel PSH, protože je jako víceoborový tezaurus samozřejmě mnohem méně podrobný v oblastech lékařství a biologie než specializovaný oborový MeSH. Bylo tedy rozhodnuto, že při mapování MeSH se bude postupovat až do takové hloubky podrobnosti, která bude pokrývat PSH, a na hesla z hlubších úrovní MeSH pak bude aplikováno mapování z nejhlouběji umístěného jim nadřazeného hesla, které bylo mapováno ručně. Takto automatizovaně přiřazená hesla budou mít označení „b“, tedy že jim bylo namapováno heslo PSH popisující nadřazený pojem. Poslední mapované heslo větve, které bude využito k rozšíření mapování na podřazená hesla, bude při mapování označeno kódem „p“ – platí i pro podřazená hesla. Rozšíření mapování na podřazená hesla je realizováno pomocí notací odrážejících umístění posledního mapovaného deskriptoru v hierarchii. Notace podřazených hesel má stejný začátek jako poslední mapované heslo, a tak je mapování snadno automatizovaně rozšířeno na všechny deskriptory s delší notací se stejným počátkem. Poslední mapování samozřejmě nesmělo být typu „n“, u kterého by nemuselo platit tvrzení, že automatizovaně přiřazené mapované heslo označovalo nadřazený pojem.

V bibliografických záznamech ovšem informace o notaci přiřazeného hesla MeSH není uvedena, jelikož každý deskriptor může být ve struktuře MeSH zařazen několikrát a na různých místech, a má tedy i různý počet notací. Bylo tedy nutné počítat s tím, že rozšíření mapování bude muset být provedeno předběžně (ne až při konverzi záznamů po sklizení pomocí protokolu OAI-PMH) a převedeno do dvojic „číslo záznamu deskriptoru MeSH“ a „identifikátor PSH“.

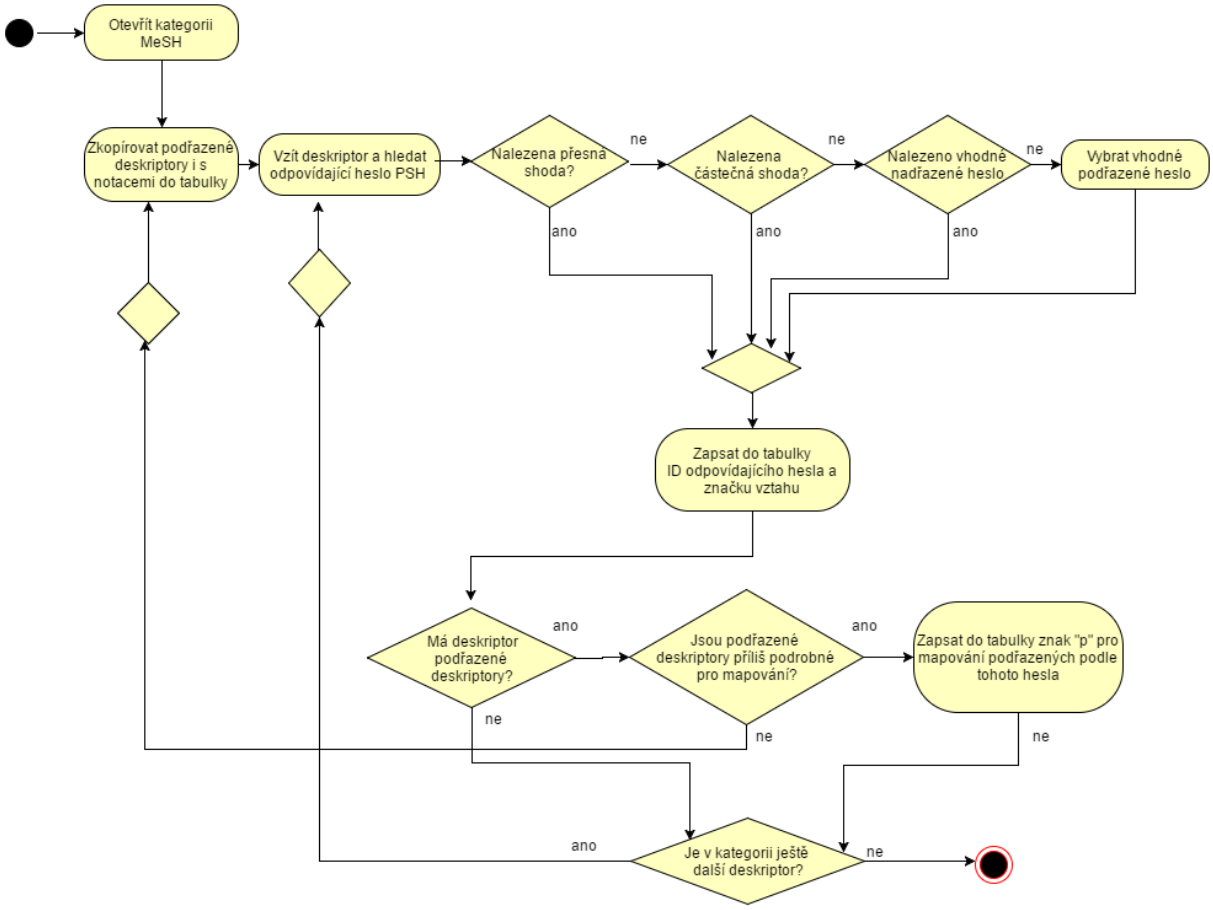
Na základě uvedených informací jsem postup mapování a dalšího zpracování stanovila následovně:

1. Proces mapování (znázorněn v diagramu aktivit na obr. 24)
    a. Ve webovém rozraní se postupuje po jednotlivých kategoriích a postupně jsou do tabulky přebírány seznamy notací a hesel ze stromu MeSH (obr. 23)

Obrázek 23: Ukázka zobrazení MeSH stromu ve webovém portálu Medvik

b. Postupně se přistupuje k jednotlivým deskriptorům a přiřazuje se odpovídající heslo PSH spolu s označením vztahu (postup je stejný jako u mapování skupin Konspektu). Pokud má deskriptor podřazená hesla, postupuje se tak hluboko, dokud se nejedná o příliš specifická hesla. Pokud jsou hesla umístěná hlouběji ve sktruktuře už příliš podrobná vzhledem k PSH, je poslednímu mapovanému deskriptoru do příslušeného sloupce zapsán kód „p“, takže mapování tohoto deskriptoru bude (v jiném vztahu) aplikováno i na podřazená hesla.

2. Další zpracování
   a. Na základě notací jsou vyhledána v XML verzi MeSH čísla deskriptorů a jsou k nim přiřazeny identifikátory namapovaných hesel PSH. Zároveň jsou vyhledány vyznačené poslední mapované deskriptory v hierarchii a všem jejich podřazeným deskriptorům jsou přiřazena mapování podle posledního ručně mapovaného deskriptoru. I tato mapování jsou převedena na dvojice „číslo záznamu deskriptoru MeSH“ a „identifikátor PSH“.
   b. Výběr jednoho vhodného mapování.
   c. Převod finálního souboru mapování do modulu Invenia a zakomponování do procesu konverze záznamů přejímaných pomocí protokolu OAI-PMH z NLK.



Obrázek 24: Diagram aktivit procesu mapování MeSH - PSH

Opakování deskriptoru ve struktuře MeSH má za následek i to, že během samotného ručního mapování, a o to více během rozšiřování mapování na podřazené deskriptory, bývá jednomu deskriptoru přiřazeno více různých hesel PSH. Norma tuto možnost připouští, pokud se samozřejmě nejedná o dvě různá mapování typu „přesná shoda“. Pokud by nebyla tato záležitost dále ošetřena, bylo by v záznamu na základě jednoho hesla MeSH přiřazeno neúměrné množství PSH hesel. Z následující tabulky (tabulka 7) je patrné, že sice téměř 49 % deskriptorů MeSH má pouze jedno mapování, ale u velkého počtu deskriptorů jsou tři a více (výjimečně až 12) mapování (celkem tvoří téměř 20 %).

Počet mapování deskriptoru	Počet záznamů deskriptorů	Poměr
1	12 940	48,961 %
2	8 336	31,541 %
3	3 474	13,145 %
4	1 071	4,052 %
5	402	1,521 %
6	136	0,515 %
7	51	0,193 %
8	11	0,042 %
9	4	0,015 %
10	2	0,008 %
11	1	0,004 %
12	1	0,004 %

Tabulka 7: Počty mapování vytvořených pro jednotlivé deskriptory MeSH

Samotný počet mapování navíc nemusí odpovídat počtu přiřazovaných hesel PSH, protože opět bylo využíváno i mapování „jeden-na-mnoho“, kdy jednomu deskriptoru MeSH byla namapována kombinace dvou hesel PSH. Při náhledu na reálný počet přiřazovaných hesel PSH mapováním je problematičnost neřešení této situace ještě zřejmější. O extrémní případy s pěti a více hesly PSH by se sice jednalo pouze ve 4 % případů, ale v bibliografických záznamech je zpravidla několik deskriptorů MeSH, což v případě aplikace mapování na bibliografické záznamy vede k násobení celého problému. Počty hesel PSH mapovaných na deskriptory jsou zobrazeny v tabulce 8.

Počet mapovaných hesel PSH	Počet záznamů deskriptorů	Podíl z celkového počtu mapování
1	12 306	46,562 %
2	8 111	30,690 %
3	3 487	13,194 %
4	1 430	5,411 %
5	524	1,983 %
6	292	1,105 %
7	123	0,465 %
8	64	0,242 %
9	35	0,132 %
10	29	0,110 %
11	12	0,045 %
12	5	0,019 %
13	5	0,019 %
14	4	0,015 %
15	1	0,004 %
20	1	0,004 %

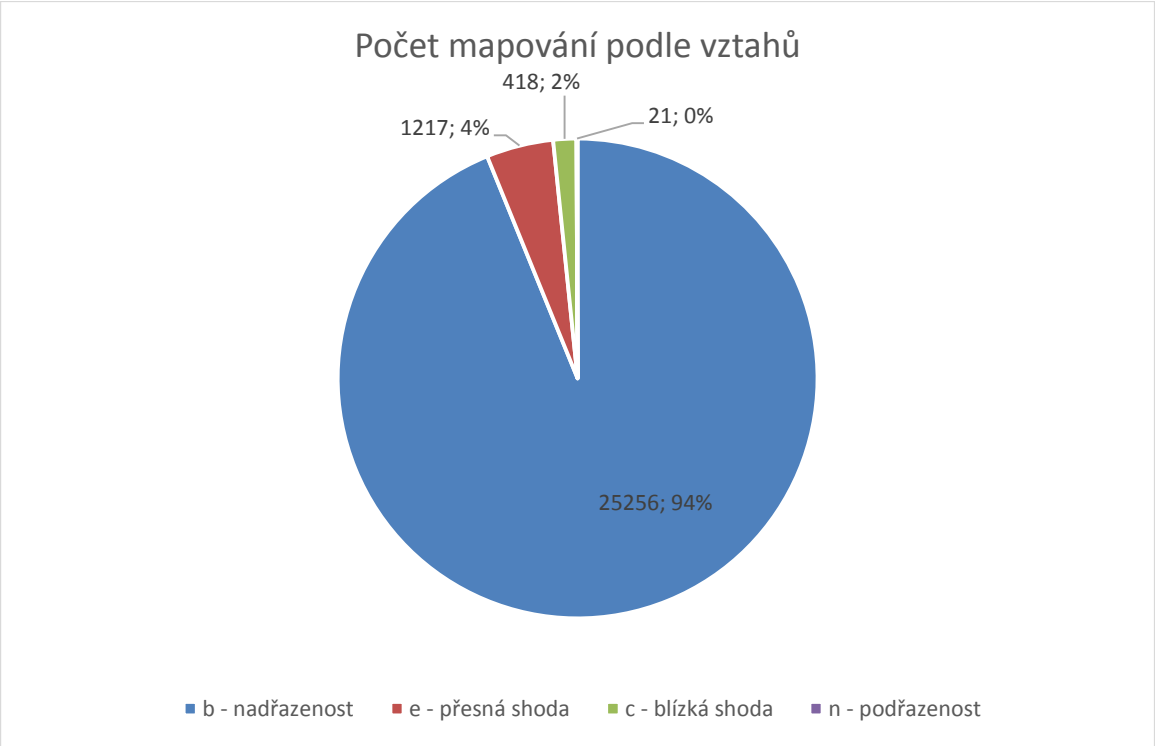
Tabulka 8: Počet hesel PSH namapovaných na deskriptory MeSH

Z mapování jednotlivých deskriptorů bylo tedy nutné vybrat to nejlepší, a to kvůli velkému množství deskriptorů opět automatizovaně. K tomuto výběru se přistupuje v první fázi na základě typu mapování podle vztahu a ve druhé na základě hierarchické úrovně mapovaného hesla PSH.

Na úvod zpracování jsem z kompletního souboru všech deskriptorů MeSH s mapováním vyřadila ty, které měly pouze jedno mapování, které nemusely být dále řešeny. Primární volbou při výběru bylo mapování typu „e – přesná shoda“, pokud měl deskriptor mapování tohoto typu, bylo vybráno a deskriptor byl vyřazen z dalšího zpracování. U tohoto typu mapování bylo při případném konfliktu, tedy přidělení více takových mapování jednomu deskriptoru, postupováno individuálně.

Obdobně bylo postupováno u mapování typu „c – blízká shoda“ a „n – podřazenost“. Typu „n“ bylo jen malé množství a většinou to bylo jediné mapování, které bylo k dispozici. Tímto postupným odebíráním zbyla množina deskriptorů MeSH (respektive jejich kódů), které měly pouze mapování typu „b“. Jednotlivým mapováním proto byla přiřazena hodnota odpovídající hloubce umístění namapovaného hesla PSH v hierarchii PSH. V případě mapování „jeden-na mnoho“ byl vytvářen průměr těchto hodnot. Pro finální verzi mapování bylo vybíráno to s nejvyšší hodnotou, tedy s nejspecifičtějším přiřazovaným heslem.

Z takto ošetřených dat vychází následující statistiky mapování MeSH-PSH: celkem bylo vytvořeno 26 912 mapování, z toho přibližně 3 600 ručně. Vzhledem ke specifickému zaměření tezauru MeSH oproti všeobecnému PSH převažuje mapování typu „b – nadřazenost“ a stejně jako v případě mapování skupin Konspektu následuje výskyt přesné shody, částečné shody a mapování typu „n“. Poměry počtů jednotlivých typů mapování jsou zobrazeny v grafu na obr. 25.



Obrázek 25: Počet mapování deskriptorů MeSH na hesla PSH podle typu vztahu

Z celkového počtu mapování jich 25 998 bylo typu „jeden-na-jeden“ a 914 „jeden-na-mnoho“ (přidělovány byly kombinace maximálně dvou hesel PSH).

Nenamapovány zůstaly tři hlavní kategorie, které nemají v PSH ekvivalent:

- [V] Publikační charakteristiky – formální charakteristika dokumentu.
- [Y] Kvalifikátory (podhesla) – upřesňují deskriptory, ale samy jsou obecné.
- [Z] Geografická místa – z PSH jsou zeměpisná označení vyloučena, není tedy možné je namapovat.

6.3.4.1. Aktualizace mapování

Jelikož každý rok vychází aktualizace českého překladu MeSH (v závislosti na aktualizacích anglické verze), je nutné aktualizovat i mapování. Kromě aktuální verze tezauru je na stránkách NLK dostupný i seznam změn mezi starou a novou verzí. Díky tomu, že mapování není závislé na slovním vyjádření termínů jednotlivých deskriptorů, není nutné řešit jejich meziroční změny. Je nutné se zaměřit pouze na nová a smazaná hesla a přesuny v hierarchii. Změny je možné snadno zapracovat do původní tabulky a pak už jen automaticky nechat zpracovat již vytvořenými nástroji.

6.3.4.2. Aplikace mapování MeSH – PSH

Mapování bylo vyzkoušeno na stejné množině záznamů jako mapování Konspekt-PSH. K popisu těchto 3 802 bibliografických záznamů bylo původně použito 5 303 jedinečných hesel MeSH, přičemž celkově bylo v záznamech použito 22 481 hesel MeSH. V tabulce 9 je zobrazen přehled 10 nejčastěji užitých hesel MeSH.

MeSH deskriptor - preferované znění	Počet užití v bibliografických záznamech
dítě	234
polymerázová řetězová reakce	179
rizikové faktory	150
prognóza	143
cytokiny	108
modely nemocí na zvířatech	108
biologické markery	101
kojenec	101
nádory	92
polymorfismus genetický	83

Tabulka 9: 10 nejčastěji užitých deskriptorů MeSH ve zkoumaných záznamech

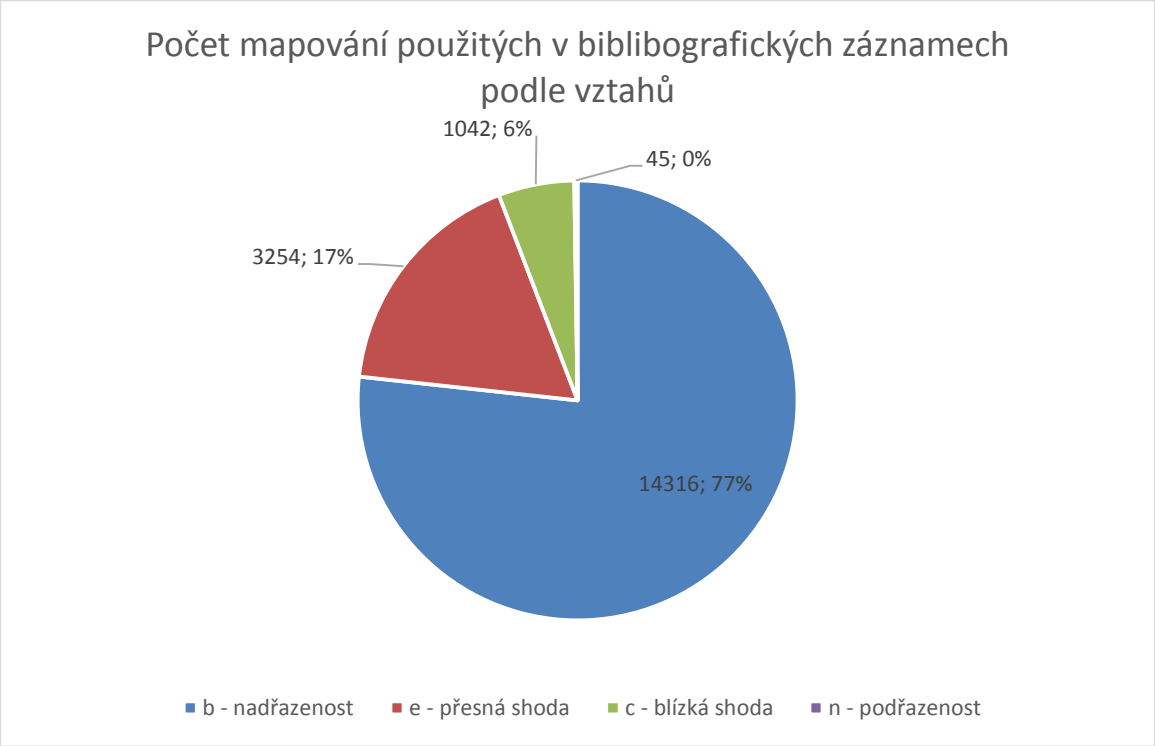
Na základě těchto hesel bylo pomocí mapování přiřazeno 760 jedinečných hesel PSH, které byly dohromady užity 18 657krát. Počet užitých hesel PSH je nižší než počet původních hesel, protože pokud bylo více hesel MeSH z jednoho bibliografického záznamu převedeno na stejné heslo PSH, došlo k deduplikaci a přiřazeno bylo jen jedno heslo. Tyto situace jsou důsledkem toho, že MeSH je oproti PSH podrobnější a rozlišuje více pojmů. Následující tabulka ukazuje 10 hesel PSH, která byla mapováním do bibliografických záznamů přidána nejčastěji.

Heslo PSH	Počet užití
výzkumné techniky	1082
epidemiologie	690
biologické faktory	482
nádory	433
bílkoviny	392
patologie	339
diagnóza	318
účinky léků	309
genetika	302
nemoci cév	254

Tabulka 10: 10 hesel PSH nejčastěji přiřazených na základě mapování MeSH

Na záznamy byly použity všechny typy mapování podle vztahů. Opět převažuje mapování typu „b – nadřazenost“, ale ne tak dramaticky, jak by bylo možné očekávat podle převahy těchto mapování v celém souboru mapování MeSH-PSH. V grafu na obr. 26 jsou uvedeny přesné

počty aplikací jednotlivých typů mapování na množině bibliografických záznamů a současně znázorněny poměry jejich četnosti.



Obrázek 26: Počty mapování podle typu vztahu, která byla aplikována na množinu záznamů

6.4. Mapování klíčových slov

Vysokoškolské kvalifikační práce tvoří přibližně 75 %[28] záznamů obsahu repozitáře NUŠL Invenio a naprostá většina z nich je přebírána z jiných systémů konkrétních vysokých škol (ručně je vloženo pouze několik závěrečných prací ze soukromé Literární akademie). Z údajů věcného popisu obsahují tyto záznamy volná autorská klíčová slova a ve třetině případů i autorský abstrakt.

Použitá volná klíčová slova nejsou obsažena v žádném slovníku a nejsou mezi nimi vyjádřeny žádné vztahy. Není tedy nijak pokryta synonymie ani hierarchické vztahy. Problematický je i fakt, že autory klíčových slov přiřazovaných k VŠKP jsou většinou sami studenti, kteří nemívají zkušenosti s indexováním dokumentů. Další práce s takovouto indexací je obtížná i v případě užití automatické indexace.

[28] Údaj platný ke květnu 2016.

Jedním z úkolů této práce bylo zjistit, zda by bylo možné využít mapování i v případě volných klíčových slov. Vycházela jsem z předpokladu, že sice asi nebude možné intelektuálně vytvořit mapování všech klíčových slov obsažených v záznamech z testovaného zdroje, ale že by mohlo postačit namapovat alespoň nejčastěji používaná klíčová slova, aby byla pokryta většina záznamů.

Před samotným mapováním bylo nutné provést výběr nejvhodnějšího zdroje záznamů VŠKP. Zdroj měl být dle požadovaných parametrů sklízen přímo do systému NUŠL Invenio (většina vysokých škol je napojena pouze na vyhledávací rozhraní NUŠL) a měl být nalezen co nejlepší poměr mezi množstvím klíčových slov a četnosti jejich užití ve všech záznamech VŠKP daného zdroje.

Zkoumány byly záznamy z České zemědělské univerzity, Vysoké školy ekonomické a Jihočeské univerzity, které byly již sklizeny do repozitáře NUŠL Invenio.

6.4.1. Analýza klíčových slov v záznamech vybraných VŠ

Pomocí skriptu v programovacím jazyce Python byly postupně hledány záznamy konkrétní vysoké školy a v každém záznamu byla nalezena klíčová slova. Pokud nově nalezené klíčové slovo ještě nebylo použito v předchozích záznamech, bylo přidáno do takto se postupně vytvářejícího seznamu (v jazyce Python šlo vlastně o slovník) a byla mu přiřazena hodnota 1. Pokud již bylo klíčové slovo na seznamu, byla pouze zvýšena hodnota o jednu. Takto vznikla tabulka obsahující veškerá klíčová slova ze záznamů dané vysoké školy včetně údaje o počtu záznamů, ve kterých byla užita.
Na základě těchto dat byly jednotlivé zdroje srovnávány v několika parametrech:

1) Počet jedinečných klíčových slov.
    a) Počet jedinečných klíčových slov použitých v množině záznamů 10 krát a víckrát.
    b) Počet jedinečných klíčových slov použitých v množině záznamů 5 krát a méně.
    c) Procentuální poměr a) a b) vzhledem k celkovému počtu klíčových slov.
2) Počet použití klíčových slov (Počet jedinečných klíčových slov vynásoben počtem záznamů, ve kterých jsou užity)
    a) Počet použití klíčových slov, která byla přiřazena v 10 a více záznamech.
    b) Počet použití klíčových slov, která byla přiřazena v 5 a více záznamech.
    c) Procentuální poměr a) a b) vzhledem k celkovému počtu použití klíčových slov.
3) Poměry počtu jedinečných klíčových slov a počtu jejich využití v záznamech.

Data jsou uvedena ve shrnující tabulce 11.

	ČZU	JČU	VŠE
Počet záznamů	19 731	24 747	42 349
Počet jedinečných klíčových slov	45 098	49 267	63 705
Počet jedinečných klíčových slov na záznam (počet záznamů/počet klíčových slov)	2,29	1,99	1,50
Celkový počet použití klíčových slov	130 851	72342	181 112
Počet použití na klíčové slovo (počet použití/počet klíčových slov)	2,90	1,47	2,84
Počet klíčových slov majících alespoň 10 opakování	1 936	1 350	2 598
Množství jedinečných klíčových slov majících 10 a více opakování (v procentech)	4,29%	2,74%	4,08%
Celkový počet použití klíčových slov s 10 a více opakováními	61 740	36 640	86 585
… procentuální vyjádření vůči celkovému počtu použití klíčových slov	47,18%	50,65%	47,81%
Počet záznamů s klíčovými slovy majícími 10 a více opakování	13 778	14 423	21 371
… procentní zastoupení takových záznamů v celé množině záznamů	69,83%	58,28%	50,46%
Počet klíčových slov majících 9 a méně opakování	43 162	47 917	61 107
Množství jedinečných klíčových slov majících 9 a méně opakování (v procentech)	96%	97%	96%
Celkový počet použití klíčových slov s 9 a méně opakováními	69 111	72 332	94 527
Počet klíčových slov majících 5 a méně opakování	41 632	46 578	59 179
Množství jedinečných klíčových slov majících 5 a méně opakování (v procentech)	92,31%	94,54%	92,90%
Celkový počet použití klíčových slov s 5 a méně opakováními	58 106	62 742	80736

Tabulka 11: Sledované údaje k záznamům VŠ

Opakované užití klíčových slov bylo sledováno v hladinách 10+, 5-9 a 5 a méně opakování, protože během prvotní analýzy bylo zjištěno, že klíčových slov s 10 a více opakováními bylo mezi 1 300 a 2 600, takže by bylo proveditelné intelektuální mapování. S klesajícím počtem opakování klíčových slov roste počet jedinečných klíčových slov a klíčová slova s pěti opakováními byla stanovena jako hranice pro demonstraci toho, jak velké množství klíčových slov je v celém souboru záznamů opakováno pouze minimálně a jejich mapování by tak nebylo efektivní.

Jihočeská univerzita

Ve zkoumané množině záznamů bylo 24 747 záznamů VŠKP z Jihočeské univerzity. Vzhledem k rozsáhlému záběru univerzity, která nabízí více než 200 oborů na 8 fakultách v různých oblastech (fakulta ekonomická, teologická, přírodovědecká, fakulta rybářství a ochrany vod a další) (Jihočeská univerzita, 2016), bylo možné očekávat, že i množství jedinečných klíčových slov bude značné a budou se v záznamech opakovat méně, než kdyby šlo o vysokou školu s užším zaměřením. Ve srovnání se záznamy ostatních dvou VŠ se tento předpoklad víceméně potvrdil. Záznamy JU mají nejnižší hodnotu průměrného počtu užití jedinečného klíčového slova (jedno klíčové slovo je použito pouze 1,47 krát) a nejnižší procentuální podíl klíčových slov majících alespoň ve všech sledovaných hladinách (10 a více opakování, 9 a méně, 5 a méně). Jistou šanci nabízela analýza podle počtu užití jedinečných klíčových slov, jelikož klíčová slova s 10 a více opakováními tvoří 51 % všech užití klíčových

slov v záznamech JU a zároveň byl i poměrně výhodný počet klíčových slov, ke kterým by bylo potřeba vytvořit mapování (pouze 1 350). Kontrolou, kolik záznamů by bylo mapováním pokryto, bylo zjištěno, že by se jednalo pouze o 58,28 % všech záznamů JU. Toto číslo navíc představuje ideální variantu, pokud by se podařilo vytvořit mapování pro všechna hesla s 10 a více opakováními. I vzhledem k tomu, že 94,5 % klíčových slov je v záznamech této univerzity opakováno pouze 5 krát a méně, nebyla klíčová slova z jejích záznamů vybrána k dalšímu zpracování.

Vysoká škola ekonomická

Vzhledem k užšímu zaměření VŠE oproti JU bylo předpokládáno, že záznamy budou obsahovat menší množství jedinečných klíčových slov v poměru k celkovému počtu záznamů a bude tedy i větší míra opakovaného užití jedinečných klíčových slov v záznamech. Ve 42 349 zkoumaných záznamech bylo nalezeno 63 705 klíčových slov, což v přepočtu znamená 1,5 klíčového slova na záznam a v tomto parametru se jedná o nejlepší výsledek mezi zkoumanými zdroji záznamů. Při vyhodnocování klíčových slov z hlediska počtu jejich užití v záznamech bylo vypočteno, že průměrně je každé klíčové slovo užito 2,84krát.

Při vyhledání nejpoužívanějších klíčových slov, která byla v celé množině zkoumaných záznamů z VŠE použita alespoň 10krát, bylo zjištěno, že v počtu 2 598 jedinečných klíčových slov tvoří pouze 4,08 % z celkového počtu klíčových slov. Z pohledu využití klíčových slov sice představují více než 47 % všech použití klíčových slov, ale i tak by po jejich namapování bylo nejméně jedním heslem PSH označeno pouze 50,46 % záznamů.

Česká zemědělská univerzita

Přestože je ČZU stejně jako VŠE zaměřena úžeji než JU a zároveň poskytla do NUŠL nejmenší množství záznamů VŠKP (19 731), neprojevily se tyto skutečnosti nijak zásadně na snížení počtu jedinečných klíčových slov. V záznamech ZČU bylo nalezeno 45 098 jedinečných klíčových slov, tedy sice nejméně ze zkoumaných zdrojů, ale nikoli výrazně. Záznamy ze ČZU si sice vedou dobře v parametrech jako počet použití klíčového slova vzhledem k celkovému počtu použití klíčových slov (průměrně bylo klíčové slovo opakováno 2,9krát) a množství klíčových slov s 10 a více opakováními v záznamech, které tvoří 4,29 % celkového počtu klíčových slov, ale celkově zde stejně jako u zbývajících zdrojů nastává problém s tím, že i při namapování nejpoužívanějších klíčových slov (s 10 a více užitími) by bylo možné indexovat hesly PSH pouze omezené množství všech záznamů zdroje. V případě ČZU by se jednalo až o 69,83 % záznamů, pokud by se podařilo namapovat na PSH všechna

klíčová slova s 10 a více opakováními. Jelikož se jedná o nejlepší výsledek mezi zkoumanými zdroji, bylo analyzováno 500 nejčastějších klíčových slov ze záznamů z ČZU, zda by bylo možné je namapovat. Klíčová slova byla pouze zběžně označována slovy ano/ne a při hrubém náhledu bylo vyhodnoceno, že 339 by jich namapovat bylo možné, byť jen na hesla označující tematicky nadřazené pojmy. U přibližně 161 by však namapování nebylo možné, protože byla příliš nejednoznačná, případně se jednalo o typ hesel, který se v PSH nevyskytuje (povolání, geografické označení apod.). Reálně by tedy bylo pokryto mnohem menší procento záznamů ČZU.

6.4.2. Vyhodnocení

Na základě získaných dat bylo rozhodnuto, že v případě VŠKP z těchto zdrojů by využití mapování nebylo efektivní. Ve všech třech případech je pouze necelých 8 % klíčových slov použito více než pětkrát. Pokud by bylo mapování využito alespoň k mapování klíčových slov, která se opakují v záznamech minimálně 10krát, bylo by v optimálním případě dosaženo přiřazení hesel PSH na základě klíčových slov maximálně k 69 % záznamů (což není reálné, protože velkou část klíčových slov by nebylo možné namapovat).

Samotné mapování klíčových slov na PSH se nezdá efektivní a v rámci této práce v tomto směru nebylo pokračováno. Do budoucna by ovšem mapování alespoň zmiňovaných klíčových slov s 10 a více opakováními mohlo být dobrým základem pro následnou automatickou indexaci, která by následně mohla být přesnější. To by ovšem muselo potvrdit či vyvrátit další testování, které je již za hranicí této práce.

7. Srovnání výsledků testovaných metod sjednocování věcného popisu

Obě testované metody sjednocování věcného popisu, tedy automatická indexace i mapování tezauru MeSH a kategorizačního schématu Konspekt, byly použity na množinu téměř 4 000 záznamů šedé literatury z Národní lékařské knihovny. Vzhledem k tomu, že při srovnávání jednotlivých metod bylo zapotřebí bližšího pohledu na záznamy, byla pro tento účel vybrána vzorová množina záznamů. Postup získání vzorku je podrobně popsán v následující podkapitole.

Srovnávání bude prováděno jak mezi výsledky automatické indexace a mapováními jednotlivých schémat, tak z hlediska dalšího využití sjednoceného věcného popisu při předávání dat do systému OpenGrey. Při předávání do tohoto systému je využíváno již existující mapování na schéma klasifikační SIGLE (popsáno v kapitole 5.2.).

7.1. Výběr vzorku

Pro srovnání bylo třeba z celkového počtu 3 802 záznamů vybrat pouze menší, zpracovatelné množství. Bylo stanoveno, že pro srovnání bude vybráno 400 záznamů vybraných podle předem stanovených kritérií. Jako primární kritérium bylo zvoleno téma dokumentu, které bylo dáno skupinou Konspektu v záznamu. Skupiny Konspektu jsou pro tento účel vhodné, jelikož je v každém záznamu vždy uvedena pouze jedna a jedná se o hrubou tematickou kategorizaci. Dalším kritériem byl rok vzniku dokumentu v závislosti na tematickém zařazení.

Nejprve byly zjišťovány požadované počty záznamů s určitými parametry, jako první spočteny záznamy označené jednotlivými skupinami Konspektu. Na základě těchto údajů byl pak učiněn přepočet, kolik záznamů z jednotlivých skupin Konspektu bude v cílovém vzorku o předpokládané velikosti 400 záznamů. Cílem bylo zachovat poměry mezi množstvím záznamů z jednotlivých tematických oblastí. V rámci tohoto přepočtu muselo dojít k zaokrouhlení, velikost výsledného vzorku byla tudíž nakonec stanovena na 397 záznamů. Zároveň byly ze vzorku vyloučeny záznamy z málo zastoupených tematických oblastí. Tyto skupiny Konspektu byly použity maximálně ve 4 záznamech a dohromady tvořily pouze 0,63 % celého objemu záznamů.

Skupina Konspektu	Počet použití v celkové množině záznamů	Kolik záznamů by bylo při zachování poměru ve vzorku 400 záznamů	Zaokrouhlený finální počet záznamů vzorku
Anatomie člověka a srovnávací anatomie	3	0,32	0
Antropologie	3	0,32	0
Biochemie. Molekulární biologie. Biofyzika	325	34,19	34
Biologické vědy	12	1,26	1
Biotechnologie. Genetické inženýrství	3	0,32	0
Buněčná biologie. Cytologie	21	2,21	2
Demografie. Populace	8	0,84	1
Ekonomie	11	1,16	1
Etika. Morální filozofie	1	0,11	0
Farmacie. Farmakologie	116	12,20	12
Filozofické systémy a hlediska	1	0,11	0
Fyzika	1	0,11	0
Fyziologie člověka a srovnávací fyziologie	191	20,09	20
Fyzioterapie. Psychoterapie. Alternativní lékařství	16	1,68	2
Geriatrie	11	1,16	1
Gynekologie. Porodnictví	78	8,21	8
Hygiena. Lidské zdraví	33	3,47	3
Chemie. Mineralogické vědy	7	0,74	1
Knihovnictví. Informatika	6	0,63	1
Lékařské vědy. Lékařství	649	68,28	68
Mikrobiologie	18	1,89	2
Obecná genetika. Obecná cytogenetika. Evoluce	38	4,00	4
Ortopedie. Chirurgie. Oftalmologie	159	16,73	17
Patologie. Klinická medicína	1713	180,22	180
Pediatrie	118	12,41	12
Přírodní vědy. Matematické vědy	1	0,11	0
Psychiatrie	55	5,79	6
Psychologie	13	1,37	1
Řízení a správa podniku	1	0,11	0
Sociální interakce	1	0,11	0
Sociální problémy vyžadující podporu a pomoc. Sociální zabezpečení	2	0,21	0
Sociologie	3	0,32	0
Stomatologie	94	9,89	10
Veřejné zdraví a hygiena	81	8,52	9
Virologie	5	0,53	1
Výchova a vzdělávání	4	0,42	0
Součet	**3802**	**400**	**397**

Tabulka 12: Počty záznamů vybraných do vzorku podle skupin Konspektu

Na základě poměrů byly určeny počty záznamů z jednotlivých skupin Konspektů (tabulka 12), které měly být ve vzorku 397 záznamů. V rámci záznamů označených stejnou skupinou Konspektu byly zjištěny roky vydání a na základě vzájemných poměrů a určeného počtu záznamů z tohoto tématu byly stanoveny počty záznamů s požadavkem splňovat obě podmínky – tedy patřičnou skupinu Konspektu a rok vydání. Pokud nebylo možné určit konkrétní požadovaný rok vydání (např. pokud byl každý záznam dané skupiny Konspektu vydán v jiném roce a ty měly tedy stejné procentuální zastoupení), byla stanovena podmínka, že může být vybrán záznam z jakéhokoliv roku z daného rozsahu.

Pro lepší představu je v tabulce 13 uveden příklad takovéhoto zpracování pro záznamy se skupinou Konspektu „Lékařské vědy. Lékařství“. Tato skupina byla v celkové množině záznamů použita 649krát a pro vzorek má být tedy vybráno 68 záznamů (viz tabulka 12). Při zpracování byly zjištěny počty záznamů podle roků vydání uvedených v záznamech označených danou skupinou Konspektu a určeny poměry těchto počtů k celkovému počtu užití této skupiny Konspektu. Poměry byly vynásobeny cílovým počtem záznamů skupiny ve vzorku před zaokrouhlením (v případě skupiny „Lékařské vědy. Lékařství“ je to 68,28). Podle předem určeného počtu záznamů, které měly být vybrány ze záznamů skupiny Konspektu (v tomto případě je cílový počet 68), je pomocí zaokrouhlení hodnot předchozího sloupce určeno odpovídající množství záznamů, které následně bude sloužit jako podmínka pro finální výběr.

Rok	Počet použití v celkové množině záznamů	Poměr záznamů z tohoto roku a celkového počtu záznamů této skupiny	Poměry násobené cílovým počtem pro vzorek o velikosti 400	Počet záznamů do vzorku z daného roku
1993	37	0,057	3,893	4,0
1992	4	0,006	0,421	0,0
1995	63	0,097	6,628	7,0
1994	72	0,111	7,575	8,0
1997	30	0,046	3,156	3,0
1996	47	0,072	4,945	5,0
1999	78	0,120	8,206	8,0
1998	59	0,091	6,207	6,0
2002	29	0,045	3,051	3,0
2003	60	0,092	6,312	6,0
2000	95	0,146	9,995	10,0
2001	61	0,094	6,418	6,0
2006	6	0,009	0,631	1,0
2007	1	0,002	0,105	0,0
2004	6	0,009	0,631	1,0
2010	1	0,002	0,105	0,0
			Součet	**68**

Tabulka 13: Příklad výběru záznamů do vzorku - skupina Lékařské vědy. Lékařství

Dle zadaných podmínek bylo z celkové množiny záznamů vybráno 397 záznamů jako vzorek ke srovnání. Závěrečný výběr podle podmínek již probíhal automaticky pomocí skriptu v jazyce Python. Výběr byl realizován z dat uložených v datovém typu slovník, jehož vlastností mimo jiné je, že data v něm uložena nejsou řazena ve stejném pořadí, v jakém byla ukládána. Při postupném výběru dat na základě daných podmínek se tím i eliminuje problém, že by se postupovalo abecedně nebo hierarchicky, a byly tak například vždy vybrány pouze nejstarší/nejnovější záznamy, případně pouze záznamy, názvy jejichž dokumentů jsou na začátku/konci abecedy. Kromě zadaných podmínek se tak do výběru nepromítají další (třeba i nechtěné) vlivy.

Potřebné údaje ze záznamů ve vzorku byly převedeny do tabulky a jsou připojeny i s popisem jako příloha na CD.

7.2. Automatická indexace

Teorie procesu automatické indexace použité na tyto záznamy byla již popsána v kapitole 5.1., nicméně zde budou zhodnoceny výsledky její aplikace na záznamy, aby mohlo dojít ke srovnání s mapováními, jejichž výsledky byly již okomentovány v předchozí kapitole.

V rámci automatické indexace byla záznamům z NLK na základě dostupných údajů přidělována hesla PSH. Automatická indexace provedená na zkoumaných záznamech mohla čerpat z následujících údajů: název, podnázev, poznámka, deskriptory MeSH a údaj o oboru NLK[29] (v záznamech označováno jako *mednas*).

Hodnocení výsledků automatické indexace bylo prováděno ručně a sledovány byly hlavně dvě skutečnosti: zda je či není heslo PSH přiřazeno zcela chybně a jestli naopak nechybí některé zásadní heslo. Pro správný věcný popis dokumentů je kromě znalosti používaných selekčních jazyků zásadní seznámení s oborem dokumentu a přístup k plnému text dokumentu. Není tedy možné, abych mohla bez přístupu k plným textům a hlavně bez odborných znalostí v daném oboru stanovovat optimální podobu věcného popisu daného záznamu a toto optimum pak srovnávala s výsledky automatické indexace (popřípadě později mapování). Mé hodnocení automatické indexace se tedy vztahuje k intelektuálně vytvořenému věcnému popisu, který je vytvářen odbornými pracovníky v NLK a je již v záznamech obsažen.

Jako chybně přiřazená hesla PSH byla označena hesla, která zásadně neodpovídala věcnému popisu a dalším údajům v záznamu. Jako chybná tak byla označena zejména hesla z neodpovídajících tematických větví PSH. Jako chybná naopak nebyla označena přiřazená hesla PSH, která sice odpovídala, ale jevila se jako příliš obecná. Na základě daných údajů se totiž nedá určit hranice, kdy je již heslo „příliš obecné“, přičemž i obecné heslo může být přínosem, pokud se jedná o tematicky odpovídající heslo.

Zda chybí „zásadní“ heslo, bylo určováno podle toho, zda jsou zastoupeny všechny oblasti, které jsou vyjádřeny v názvu a věcném popisu v záznamu. Například pokud jsou v záznamu dokumentu „Komplexní stomatologická péče o kandidáty transplantací srdce, ledvin a jater a pacienty po transplantacích“ přiřazena hesla PSH stomatologická péče (PSH12851) a zubní lékařství (PSH12850), nelze říci, že by přiřazená hesla byla chybná, ale rozhodně tu chybí zásadní informace o relevanci k transplantacím orgánů, která je v původním věcném popisu uvedena.

Jako problematické bylo alespoň jedno automaticky přiřazené heslo vyhodnoceno u 59 záznamů (tedy 14,86 % záznamů z výběru). Z toho se ve 14 případech jednalo o přiřazení hesla z tematické větve „Obecnosti“, která v PSH plní úlohu kategorie „různé“ a obsahuje řadu

[29] Jedná se p údaj z autoritního souboru oborů vytvářeného NLK, který je někdy označován i termínem „předmětové obory NLK“.

obecných širokých pojmů (například analýza, látky, parametry apod.). Tuto větev by bylo vhodnější z automatické indexace vyloučit. Ve zbývajících 49 případech bylo problémem přiřazení hesla PSH z naprosto nesouvisející větve. K tomu docházelo u obecně znějících hesel, např. monitoring (v PSH pod ekonomií), výkony (v PSH také pod ekonomií), čas (v PSH v tematické větvi fyzika), řetězové reakce (v PSH pod jadernou fyzikou, v záznamech NLK se často vyskytuje slovní spojení „řetězová reakce“, na jejichž základě bylo toto heslo systémem přidělováno, přestože významově nesouvisí s jadernou řetězovou reakcí). Pouze ve 4 případech byla chybná obě hesla a jednou šlo o kombinaci chybného hesla a nevhodného hesla z větve „Obecnosti“.

Jako chybná byla vyhodnocena i hesla PSH přidělená na základě formálních znaků dokumentu, které byly zmíněny například v názvu. Jednalo se o hesla jako „výzkum“ nebo „primární články“.

Celkově byl počet problematických automaticky přiřazených hesel PSH výrazně nižší, než bylo očekáváno (vzhledem k prováděným kontrolám popsaným v kapitole 5.1.2.1, při nichž byly prováděny úpravy téměř ve třetině záznamů). Pro tuto skutečnost se nabízí několik vysvětlení, která nebylo možné v rámci této práce ověřovat, ale do budoucna by bylo vhodné je prověřit. Jedním z důvodů může být tematická oblast lékařství a příbuzných oborů, z nichž pocházely zkoumané záznamy. Otázkou je, do jaké míry výsledky ovlivnila terminologie, která bude pravděpodobně ustálenější než například v humanitních či společenských vědách. Druhým důvodem, ke kterému se přikláním (s tím, že svou roli mohl sehrát i první uvedený důvod), je, že tak kvalitní automatická indexace je výsledkem již dobře provedené intelektuální indexace, již provedli odborní pracovníci v NLK. Pro zjištění převažujícího faktoru by bylo vhodné porovnat výsledky automatické indexace kupříkladu s automatickou indexací VŠKP z lékařských fakult (u VŠKP často tvoří klíčová slova sami studenti). Záznamy většího množství takových prací však nejsou v době psaní této práce k dispozici.

Při pohledu na automatickou indexaci vybraných záznamů z perspektivy (ne)chybějícího zásadního hesla PSH pro popis dokumentu již výsledky tak dobré nejsou. Vyhodnocovat, zda heslo chybí, je samozřejmě obtížnější než hodnotit relevanci přiřazeného hesla. Jak bylo již uvedeno, výchozími body byl již existující věcný popis vytvořený v NLK. Na základě deskriptorů MeSH, oborů NLK a případně názvu byly určeny oblasti, které by měly být pokryty hesly PSH s tím, že se samozřejmě nepočítá s přidělením stejného počtu hesel, jelikož PSH

nezachází v oblasti lékařství a příbuzných oborů do takové hloubky. Uvádím příklad určení, že v záznamu po automatické indexaci chybí heslo PSH z nějaké oblasti:

Id záznamu: oai:medvik.cz:114139
Název: Analýza nákladů a užitků péče u pacientů s chronickým selháním ledvin
Obory NLK: nefrologie; management, organizace a řízení zdravotnictví; ekonomie, ekonomika, ekonomika zdravotnictví
Deskriptory MeSH: hemodialýza; zdravotní péče - náklady; chronické selhání ledvin; náklady a analýza nákladů
Automaticky přiřazená hesla PSH: náklady (PSH1780); péče o pacienta (PSH13080)

Jak je vidět, automatická indexace dobře podchytila, že se jedná o záznam popisující dokument věnující se nákladům na péči o pacienta, ale chybí heslo PSH ukazující souvislost s nemocemi ledvin nebo nefrologií. Obdobně to platilo i v ostatních záznamech, u nichž bylo označeno, že mezi jejich automaticky přiřazenými hesly PSH chybí pokrytí nějaké oblasti. Celkově bylo ve výběru 234 takových záznamů. Část těchto případů je dána limitem maximálně 2 automaticky přiřazovaných hesel PSH na záznam.

7.3. Srovnání automatické indexace a mapování skupin Konspektu

Srovnávána byla hesla PSH přiřazená záznamům ze vzorku pomocí automatické indexace a skrze mapování Konspektu. Zásadním rozdílem mezi výsledky obou přístupů je množství přiřazených hesel. Zatímco pomocí automatické indexace jsou přiřazována dvě hesla PSH (každý záznam ve vzorku obsahoval dostatek údajů k přiřazení dvou hesel), pomocí mapování skupin Konspektu bylo přiřazeno vždy jen jedno heslo PSH.

Skupiny Konspektu představují pouze hrubou kategorizaci, což se odráží i ve výsledcích jeho mapování na PSH. Záznamům vzorku bylo na základě mapování skupin Konspektu přiřazeno pouze 18 různých hesel PSH. Oproti tomu bylo výsledkem automatické indexace přidělení 328 různých hesel PSH. Část tohoto nepoměru je samozřejmě důsledkem rozdílného množství přidělovaných hesel na základě obou postupů, ale zásadní je obecnost zdrojového mapovaného slovníku ve srovnání s metodou automatické indexace vybírající k popisu co nejpodrobnější hesla PSH.

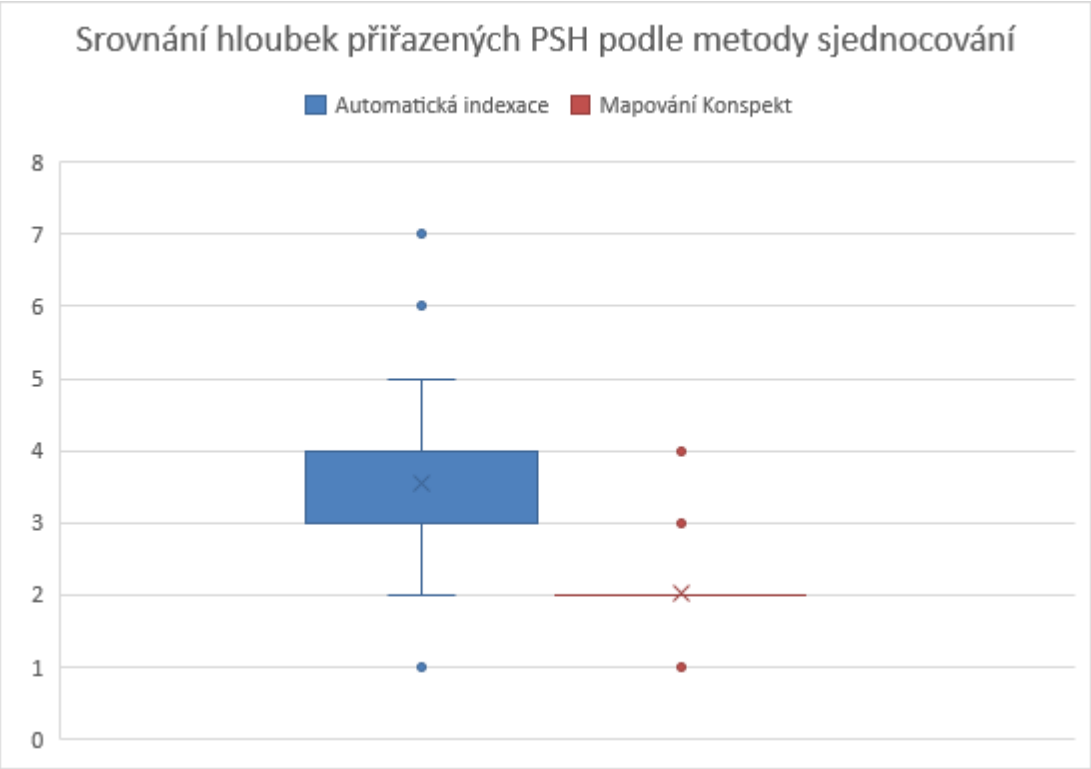
Podrobnost přiřazených hesel lze srovnat pomocí již představeného přístupu, kdy je počítáno s hloubkou umístění hesla v hierarchii hesláře PSH. Hlavní hesla mají hodnotu 1 a každá další úroveň hloubky je označena číslem o 1 větším.

Tabulka 14 ukazuje počty hesel PSH o dané hodnotě hloubky umístění v hierarchii, která byla přiřazena pomocí obou srovnávaných metod do záznamů vzorku.

	Způsob přiřazení hesel PSH	
Hloubka hesel PSH	Automatická indexace	Mapování Konspekt
1	14	37
2	104	320
3	269	39
4	287	1
5	90	0
6	29	0
7	1	0
Součet	794	397

Tabulka 14: Počty hesel PSH dané hloubky, která byla přiřazena do záznamů vzorku

Pomocí automatické indexace je přiděleno dvojnásobné množství hesel, ale i tak lze sledovat, že hesla z mapování Konspektu jsou obecnější a celkově téměř nezasahují do hlubších úrovní PSH. Tento trend ještě lépe vynikne při zobrazení těchto dat v krabicovém grafu na obr. 27, který umožňuje srovnat centrální tendence dat. Krabice představuje seřazené hodnoty mezi 1. a 3. kvartilem a obsahuje tedy 50 % dat. Z hran krabice vycházejí antény (někdy také označovány jako vousy nebo tykadla) označující hranici po přičtení nebo odečtení 1,5 hodnoty kvartilu (u spodní antény se odečítá hodnota 1,5 násobku 1. kvartilu a u horní antény se přičítá hodnota 1,5 násobku 3. kvartilu) (Hendl, 2015, s. 107). Tečky pak označují výrazně odchýlené hodnoty, které se označují jako „podezřelé“. Graf na obr. 27 ukazuje, že v případě automatické indexace je 50 % hesel o hloubce 3 a 4, hesla o hodnotě hloubky 2 a 5 se vyskytují ještě v přiměřeném množství, ale hlubší i obecnější hesla jsou spíše výjimečná. U hesel přiřazených podle mapování Konspektu tak výrazně převažuje hodnota hloubky 2, tedy že není možné zkonstruovat krabici ani antény, protože hesla jiných hodnot jsou touto metodou přiřazena velice zřídka.



Obrázek 27: Srovnání hloubek přiřazených PSH pomocí automatické indexace a mapování Konspektu

Různý počet přiřazených hesel PSH byl překážkou srovnání sjednocení věcného popisu. Pro výpočet jedné definitivní hodnoty hloubky pro záznam byl z hesel přidělených do záznamu pomocí automatické indexace vypočten průměr hloubek obou hesel PSH. Dostaneme tak jakousi průměrnou hodnotu hloubky věcného popisu pomocí PSH pro záznam. Tabulka 15 ukazuje počty záznamů s těmito jednotlivými hodnotami.

	Počet záznamů podle hloubek průměru přiřazených hesel	
Hloubka	Automatická indexace	Mapování Konspektu
1	0	37
1,5	2	0
2	14	320
2,5	52	0
3	68	39
3,5	98	0
4	98	1
4,5	45	0
5	17	0
5,5	3	0

Tabulka 15: Počty záznamů s jednotlivými hodnotami hloubek hesel nebo kombinací hesel PSH

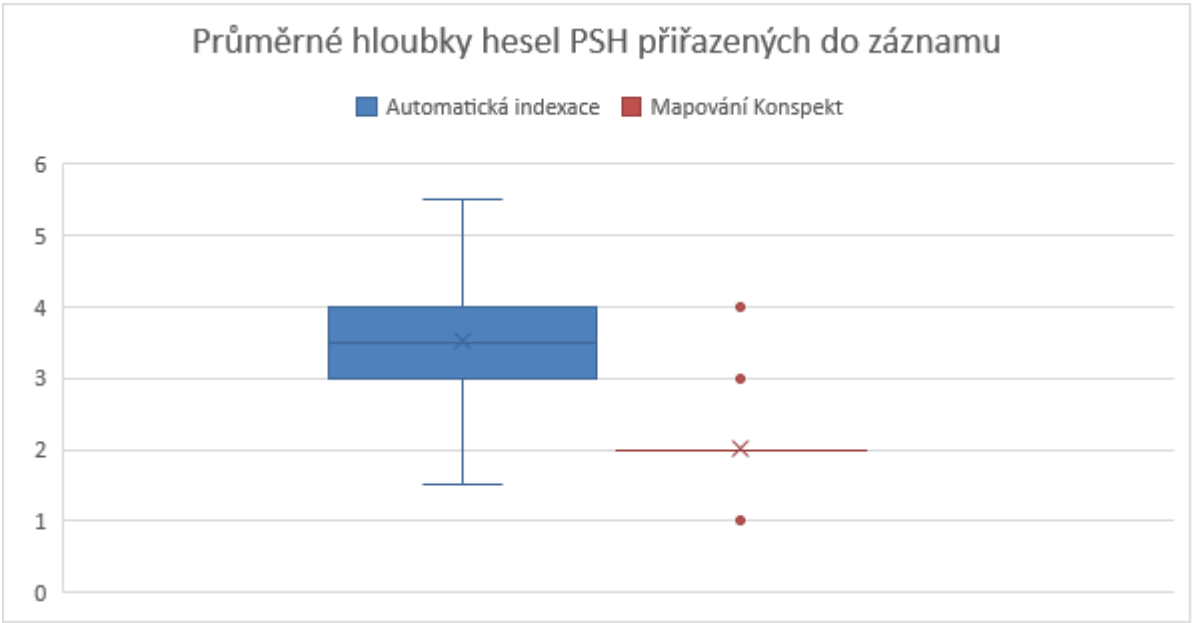
Údaje z tabulky 15 potvrzují, že kombinace hesel přiřazená pomocí automatické indexace je podrobnější než obecná hesla PSH přiřazená na základě mapování skupin Konspektu. Z hodnot

pro jednotlivé záznamy byl vypočten průměr, modus (nejčastější hodnota) a medián (střední hodnota) hloubek hierarchického umístění hesel PSH pro množinu záznamů ve vzorovém souboru (viz tab. 16), které tento předpoklad dále potvrzují. Vypočten byl i rozptyl hodnot výběru a směrodatná odchylka výběru, která určuje, jak výrazně jsou hodnoty rozptýleny či odchýleny od průměru hodnot.

	Automatická indexace	Mapování Konspektu
Průměr	3,54	2,01
Medián	3,5	2
Modus	4	2
Rozptyl	0,58	0,20
Směrodatná odchylka	0,76	0,45

Tabulka 16: Vlastnosti hodnot hloubek hesel PSH přiřazených do záznamů pomocí automatické indexace a mapování Konspektu

Hlavní výhodou zprůměrování je, že je možné porovnávat stejný počet hodnot. Výsledky vyhodnocení průměrů hloubek hesel PSH v záznamech (viz graf na obr. 28) víceméně odpovídají nezprůměrovaným hodnotám. V případě mapování Konspektu jde o stejná data a tedy i výsledné zobrazení. V případě zprůměrování hodnot u výsledků automatické indexace dojde k protažení antén a vymizení výrazně se odchylujících hodnot. Veškeré hodnoty se v tomto případě vejdou do rozmezí daného 1,5 násobku 1. a 3. kvartilu.



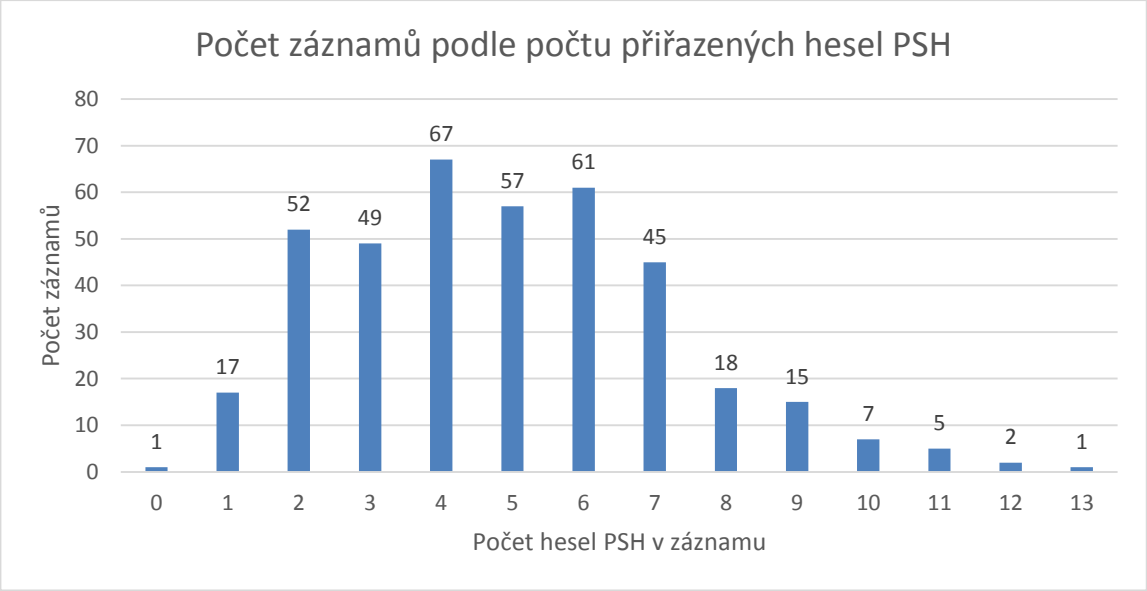
Obrázek 28: Průměrné hloubky hesel PSH přiřazených do záznamů pomocí automatické indexace a mapování Konspektu

Z pohledu přesnosti přiřazených hesel PSH platí analýza výsledků automatické indexace ukazující chybné heslo u téměř 15 % záznamů. Oproti tomu hesla přiřazená na základě

mapování Konspektu jsou vždy odpovídající, ovšem jde o velmi hrubou kategorizaci a počet přiřazených hesel je poloviční.

7.4. Srovnání automatické indexace a mapování tezauru MeSH

I v případě srovnávání výstupů automatické indexace a mapování tezauru MeSH je podstatný počet přiřazených hesel PSH. Jak již bylo zmíněno, pokud záznam obsahuje dostatek údajů k jejich přiřazení, jsou pomocí automatické indexace přidělována dvě hesla PSH, což platilo u všech záznamů vybraného vzorku. Počet hesel PSH přiřazených na základě mapování tezauru MeSH kolísá podle množství deskriptorů MeSH v záznamu. Může být vyšší, pokud je jednomu deskriptoru MeSH namapována kombinace hesel PSH, nebo naopak nižší, odpovídá-li více deskriptorům MeSH v záznamu jedno heslo PSH, které je pak přiřazeno pouze jednou. Následující graf na obr. 29 ukazuje počet záznamů s danými počty hesel PSH přiřazenými na základě mapování tezauru MeSH.



Obrázek 29: Počet záznamů podle počtu přiřazených hesel PSH pomocí mapování tezauru MeSH

V minimu případů (4,5 %) je pomocí mapování přiřazeno méně nebo stejné množství hesel PSH, jako při automatické indexaci. Ve 13,1 % záznamů je pomocí mapování přiřazeno stejné množství hesel jako při automatické indexaci a v 82,4 % je přiřazeno více hesel PSH než pomocí automatické indexace.

Přestože se ve vzorku vyskytoval pouze jeden případ záznamu bez popisu pomocí deskriptorů MeSH a v celém souboru se jedná o dva případy, je poměrně problematické, že není možné se na mapování plně spolehnout. V případě těchto záznamů byly místo deskriptorů

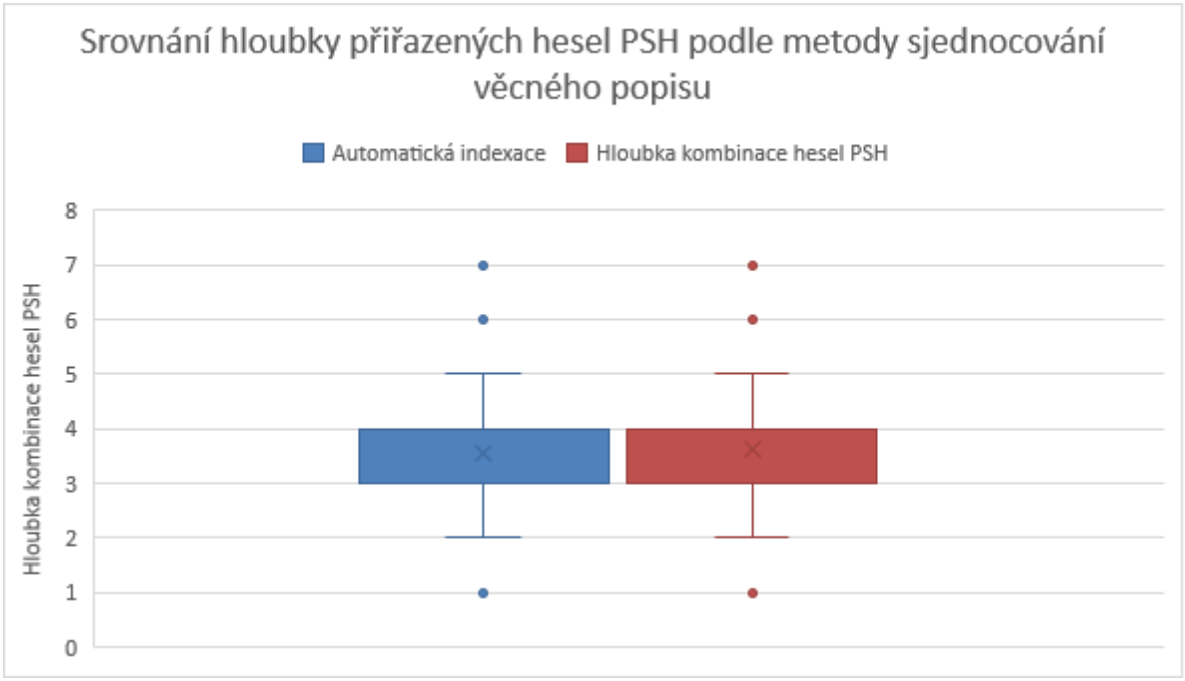
MeSH použity neautoritní formy předmětových hesel. V běžném provozu by bylo možné použít na tyto mapováním nezpracované záznamy automatickou indexaci ve chvíli, kdy by se prováděla na ostatních záznamech bez hesel PSH.

Druhým důležitým aspektem po množství přiřazených hesel PSH je jejich podrobnost, která je opět hodnocena podle hloubky umístění hesel v hierarchii PSH. V tabulce 17 jsou zobrazeny počty hesel PSH v závislosti na jejich hodnotě hloubky a způsobu, jakým byla přiřazena do záznamů.

	Způsob přiřazení hesel PSH	
Hloubka hesel PSH	Automatická indexace	Mapování MeSH
1	14	21
2	104	172
3	269	717
4	287	774
5	90	139
6	29	119
7	1	1
Součet	794	1943

Tabulka 17: Srovnání počtů hesel PSH o dané hloubce, která byla přiřazena na základě automatické indexace a mapování MeSH

Při vizualizaci dat v krabicovém grafu (obr. 30) je zjevná velká podobnost mezi výsledky, jež se liší pouze křížkem vyznačujícím průměrnou hodnotu dat, která je v případě mapování MeSH nepatrně vyšší.



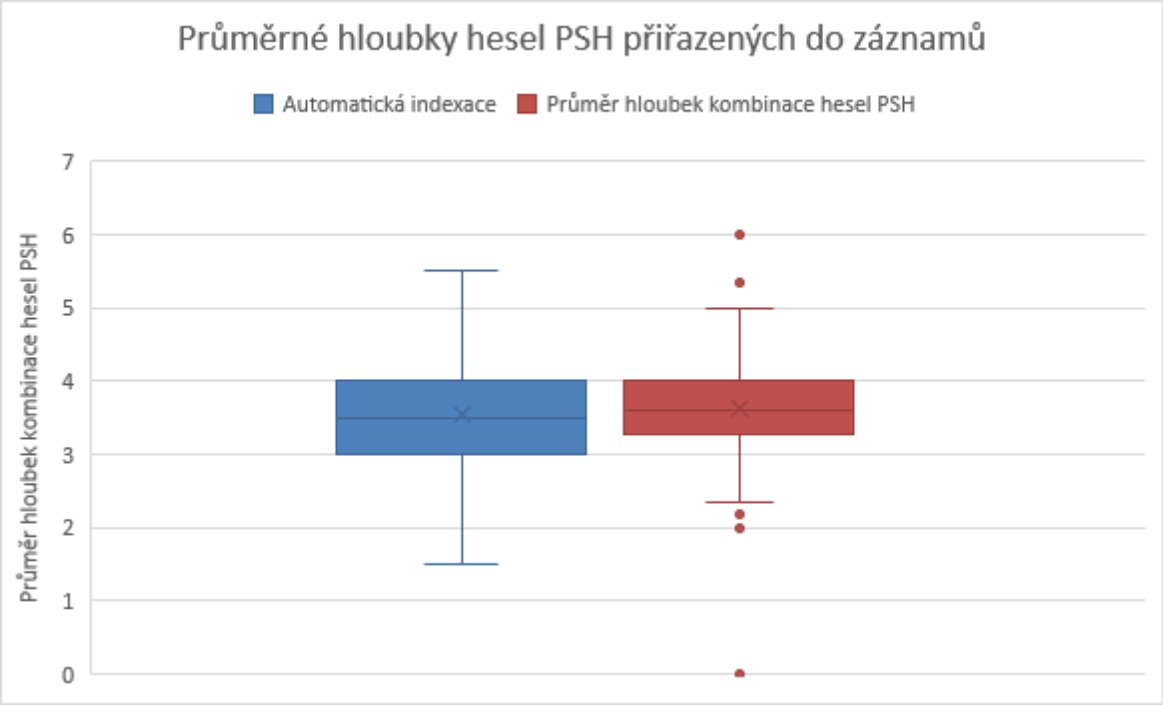
Obrázek 30: Srovnání hloubek hesel PSH přiřazených pomocí automatické indexace a mapování MeSH

Pro lepší srovnání je pro každý záznam opět vypočtena průměrná hodnota hloubky věcného popisu pomocí hesel PSH, a to jako průměr hodnot hierarchické hloubky přidělených hesel PSH. Na základě hodnot pro jednotlivé záznamy byl následně vypočten průměr, modus a medián pro celý soubor záznamů ve vzorku. Hodnoty jsou vyobrazeny v tabulce číslo 18.

	Automatická indexace	Mapování tezauru MeSH
Průměr	3,54	3,63
Median	3,5	3,6
Modus	4	4
Rozptyl	0,58	0,38
Směrodatná odchylka	0,76	0,62

Tabulka 18: Vlastnosti hodnot hloubek hesel PSH přiřazených do záznamů pomocí automatické indexace a mapování MeSH

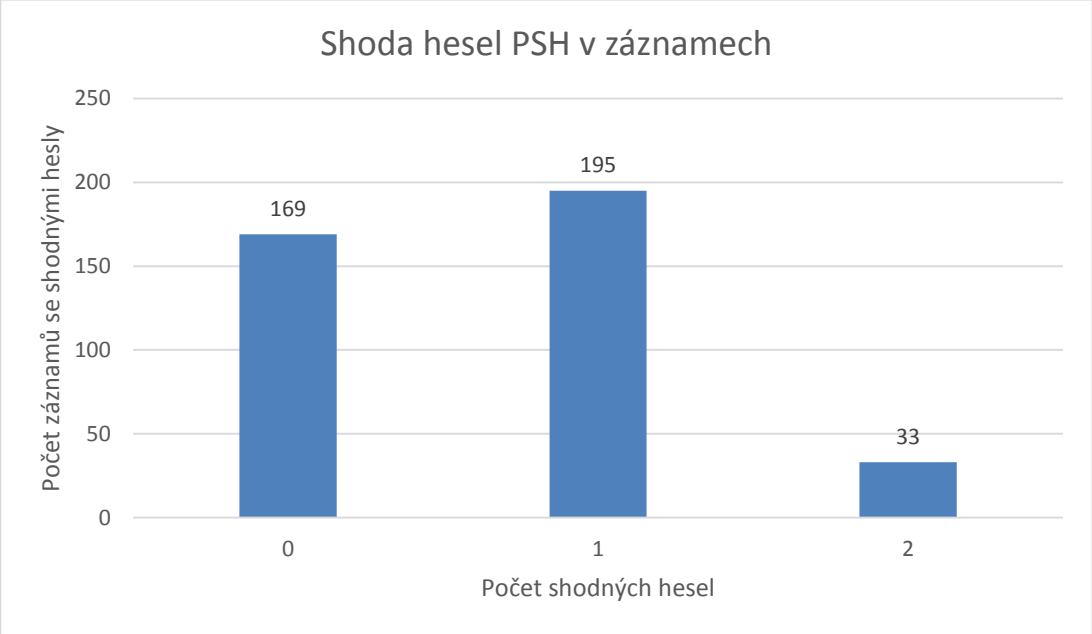
Srovnání průměrných hodnot je vyobrazeno v krabicovém grafu na obr. 31. Zde již lze pozorovat rozdíly mezi výsledky obou postupů. Polovina průměrných hloubek hesel PSH přiřazených do záznamů mapováním MeSH leží mezi hodnotami 3,27 a 4, je zde tedy patrný příklon k hodnotě 4 oproti většímu rozpětí mezi hodnotami 3 a 4 v případě automatické indexace.



Obrázek 31: Průměrné hloubky hesel PSH přiřazených do záznamu pomocí automatické indexace a mapování MeSH

V případě srovnávání výstupů obou postupů sjednocování věcného popisu je zajímavá i statistika, jak často byla přiřazena shodná hesla pomocí automatické indexace shodná s hesly přidělenými na základě mapování MeSH (obr. 32). Ta ukazuje, že v 57,4 % případů je přiděleno

pomocí obou přístupů 1-2 stejné heslo PSH. Ve 42,6 % pak není zaznamenaná přímá shoda v přidělených heslech PSH.



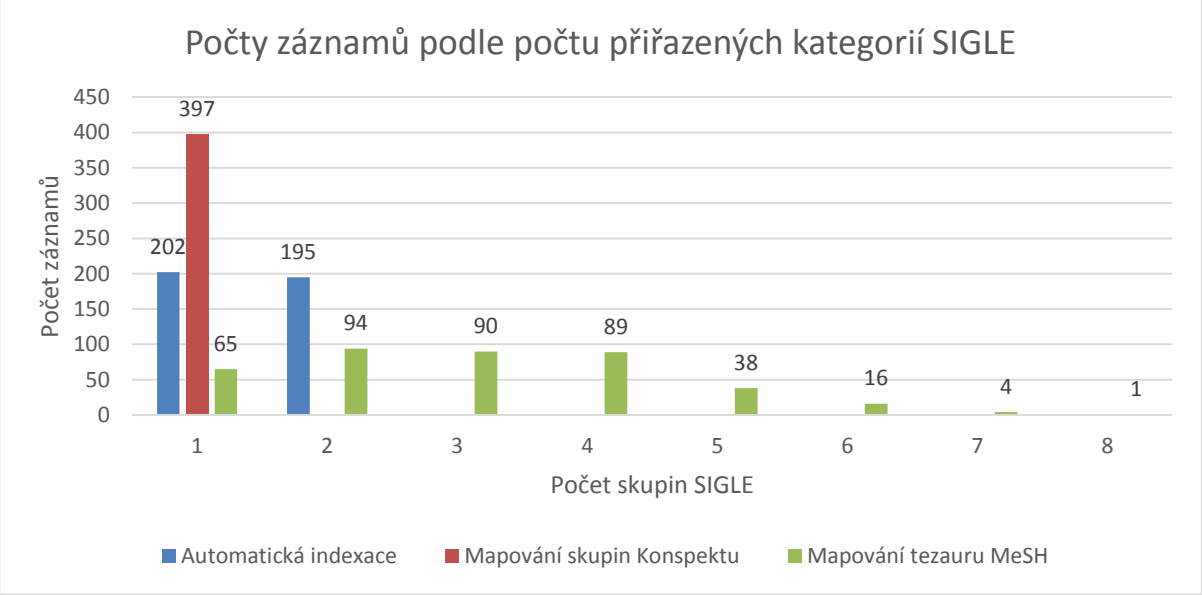
Obrázek 32: Počet shodných hesel PSH v záznamu, která byla přiřazena pomocí automatické indexace a mapování MeSH

Z pohledu přesnosti a podrobnosti je mapování MeSH sice o něco lepší volbou, ale rozdíly nejsou nijak výrazné a výsledná výhoda tohoto postupu se tak nemusí úplně vyrovnat požadavkům na vytvoření a údržbu mapování. Zásadní je ovšem rozdíl v množství přiřazených hesel a i celková spolehlivost metod. V případě mapování sice může dojít k výpadku, pokud není v původním záznamu nalezen deskriptor MeSH, nicméně právě na takovéto záznamy může být dodatečně aplikována automatická indexace. Pokud naopak dojde k chybě v automatické indexaci (a je přiřazeno nevhodné heslo), neexistuje jednoduché automatizované řešení.

7.5. Srovnání výsledků automatické indexace a mapování při úpravě záznamů pro předávání do systému OpenGrey

V kapitole 5.2 jsou popsány důvody i způsob práce se sjednoceným věcným popisem záznamů pro předávání do systému OpenGrey. Tento systém vyžaduje, aby každý záznam měl minimálně jednu kategorii z klasifikace SIGLE. Pro tento účel bylo již v minulosti vytvořeno mapování mezi PSH a klasifikací SIGLE, které zajišťuje splnění této podmínky pro předávku dat. Jedná se o velice hrubé mapování, jelikož je vytvořeno pouze do druhé úrovně hierarchie schématu SIGLE. Dochází tak k zobecnění hesel a stírání rozdílů mezi různě podrobnými hesly.

Na základě různých technik sjednocování věcného popisu je přiřazováno různé množství kategorií SIGLE. Pokud byl věcný popis sjednocen na PSH pomocí automatické indexace, byly přidělovány 1-2 kategorie SIGLE, při použití mapování tezauru MeSH 1-8 hesel a při mapování skupin Konspektu vždy pouze 1 heslo. Rozložení množství kategorií SIGLE v záznamech je vyobrazeno v grafu na obr. 33.



Obrázek 33: Počty záznamů, do kterých je přiřazeno dané množství kategorií SIGLE

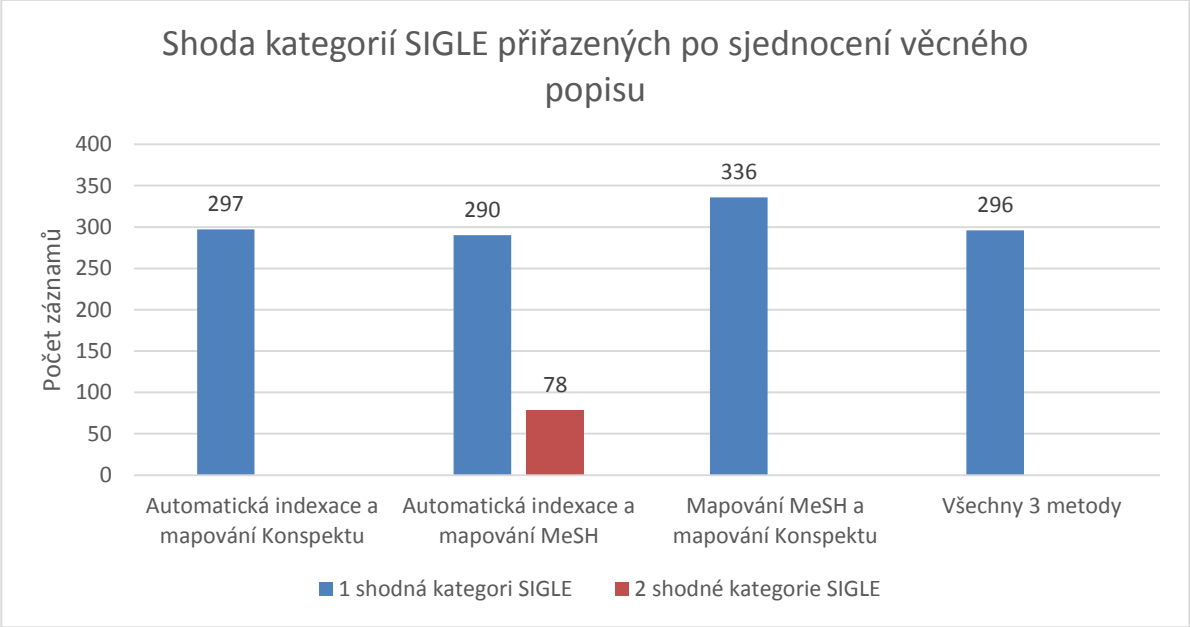
Z pohledu kvantity vyšlo jako nejlepší mapování tezauru MeSH, neboť na základě hesel PSH přiřazených touto metodou bylo následně do 83,63 % záznamů přiděleno 2 a více kategorií SIGLE. Po automatické indexaci byla pouze ve 49,11 % záznamů přidělena 2 hesla a ve zbylých pouze jedno. Na základě hesel PSH dodaných podle mapování Konspektu byla do záznamů přiřazena vždy pouze jedna skupina. Kromě nejvyššího počtu přiřazených kategorií SIGLE drží mapování MeSH prvenství i v počtu přiřazených jedinečných kategorií SIGLE (tab. 19). Záznamy by se tak v systému dostaly do více kategorií a teoreticky pak mohou být lépe a častěji vyhledané.

Způsob sjednocení věcného popisu	Počet jedinečných kategorií SIGLE	Celkový počet přiřazených kategorií SIGLE
Automatická indexace	35	592
Mapování MeSH	38	1 201
Mapování Konspekt	11	397

Tabulka 19: Počty jedinečných kategorií SIGLE a celkový počet přiřazených kategorií SIGLE

Při vyhodnocování kategorií SIGLE dodaných do záznamů bylo vypozorováno velké množství případů, kdy byla po použití různých metod sjednocení věcného popisu nakonec přiřazena shodná kategorie SIGLE. Jak ukazuje graf na obr. 34, ve 296 případech bylo jedno shodné heslo přiřazeno po použití všech metod. Další sloupce ukazují počty záznamů, do nichž

byly po použití různých metod sjednocování přiřazeny až dvě shodné kategorie SIGLE. V 78 případech byly po sjednocení věcného popisu pomocí automatické indexace použity stejné kategorie SIGLE jako po mapování MeSH.



Obrázek 34: Shoda přiřazených kategorií SIGLE pomocí jednotlivých metod sjednocování věcného popisu

Z výsledků tohoto srovnání vychází, že pro pouhé splnění podmínky pro předávání dat do OpenGrey, tedy aby každý záznam obsahoval jednu kategorii SIGLE, by postačilo mapování Konspektu, které zajišťuje přidělení právě jednoho hesla PSH a v důsledku toho i kategorie SIGLE. Při využití automatické indexace by bylo přiřazeno dvojnásobné množství kategorií SIGLE a teoreticky by tak mohla být zvýšena vyhledatelnost záznamu v systému OpenGrey, avšak za cenu několika chybně přiřazených skupin. Využití mapování tezauru MeSH by bylo z hlediska počtu přiřazovaných kategorií SIGLE nejlepší variantou, ale vzhledem k nárokům na jeho vytvoření a údržbu by v takovém případě bylo vhodné do větší hloubky propracovat i mapování mezi PSH a SIGLE, aby byl využit plný potenciál těchto propojení.

8. Závěr

V rámci této diplomové práce byly srovnávány metody sjednocování věcného popisu v záznamech agregovaných do repozitáře NUŠL. Byly představeny metody a postupy aplikované v zahraničních systémech BASE a LASSO a současné řešení tohoto problému v repozitáři NUŠL.

Na základě zkušeností z ostatních repozitářů i samotného systému NUŠL bylo navrženo srovnání automatické indexace, v NUŠL již prováděné, a mapování klasifikačních a kategorizačních schémat. Obě tyto metody měly za cíl přiřadit do záznamů hesla PSH. V rámci experimentu bylo vytvořeno mapování skupin Konspektu a tezauru MeSH. Obě mapování a automatická indexace byly aplikovány na množinu záznamů z Národní lékařské knihovny a výsledky metod byly srovnávány na vybraném vzorku tak, aby bylo možné vyhodnocovat i správnost přiřazených hesel PSH (zvláště v případě automatické indexace).

Automatická indexace se na zkoumaném vzorku ukázala jako překvapivě spolehlivá, přestože se dá spekulovat, do jaké míry je to ovlivněno poměrně stabilizovanou terminologií v oblasti medicíny a tím, že zkoumané záznamy byly vytvářeny profesionálními informačními pracovníky. Množství takto přiřazených hesel je sice dvojnásobné oproti heslům z mapování Konspektu, ale bývá jich do záznamu přiřazeno méně než pomocí mapování MeSH. Jako chybná hesla byla vyhodnocena zejména hesla z úplně nesouvisejících tematických větví PSH, takže při zobecnění (například pro předávání do OpenGrey nebo pro vytvoření hrubé předmětové kategorizace) je záznam zařazen i do naprosto nevhodných kategorií. Hlavní nevýhodou hesel přiřazených pomocí automatické indexace je ovšem časté nepodchycení všech témat v dokumentu záznamu.

Hesla z mapování MeSH jsou podrobná a jsou přiřazována ve velkých počtech. Hlavní výhodou ovšem je, že nedochází k nepokrytí části témat. V každém případě jsou totiž veškeré deskriptory MeSH reflektovány v heslu PSH, byť může dojít k zobecnění. Nevýhodou je kromě vyšších nároků na vypracování mapování i nutnost jeho každoroční údržby a nutnost část mapování vytvářet automaticky, což může vést k oslabení silných stránek této metody. Velkou nevýhodou je také závislost na věcném popisu záznamu. Zatímco v případě automatické indexace jsou hesla PSH vybrána na základě obsahu několika polí, v případě mapování MeSH se jedná jen o pole s deskriptory MeSH. Pokud je však přiřazeno nesprávné heslo nebo není-li

přiřazeno žádné, neexistuje způsob jak pomocí tohoto mapování přiřadit heslo PSH. V takovém případě musí být nasazena automatická indexace.

Výhodou hesel PSH získaných z mapování Konspektu je spolehlivost. Skupina Konspektu byla v záznamu uvedena vždy a bez chyb. Nevýhodou je samozřejmě obecnost skupin Konspektu a tedy i namapovaných hesel PSH. Ve zkoumaném vzorku nebyl problém s nepodchycením nějakého tématu pomocí hesel PSH, jelikož se jednalo o záznamy dokumentů z poměrně jasně definovaného oboru a nebyl tedy ani problém s přiřazením pouze jedné skupiny Konspektu. V případě záznamů interdisciplinárních dokumentů by mohl vzniknout problém, jelikož k popisu smí být použita pouze jedna tematická skupina Konspektu.

Na základě předložených dat jsem došla k závěru, že pro výběr vhodné techniky sjednocení dokumentů je důležité předem stanovit, za jakým účelem ke sjednocování dochází. Pokud je cílem hrubá kategorizace pro vytvoření filtru vyhledávání nebo jako v případě systému NUŠL k předávání dat do OpenGrey, postačí mapování Konspektu (kombinace využití tohoto mapování s automatickou indexací by ovšem neměla být problém).

Mapování MeSH je naopak vhodné, pokud by mělo být sjednocení věcného popisu provedeno za účelem umožnění vyhledávání pomocí hesel PSH s využitím vztahů mezi hesly způsobem, jakým jsou v odborných databázích využívány tezaury. V takovém případě je žádoucí mít v záznamu co možná nejvíce hesel PSH a zároveň roste i tlak na to, aby byla přiřazenými hesly pokryta skutečně všechna témata, jimiž se dokument záznamu zabývá. To může z testovaných metod zajistit nejlépe právě mapování podrobného schématu, jakým je MeSH.

Vzhledem k současným plánům rozvoje systému NUŠL, které zahrnují i vytvoření tematického filtru na základě přiřazených hesel PSH, bych tedy doporučovala zaměřit se na možnosti využití mapování Konspektu a případně dalších hrubších kategorizačních a klasifikačních schémat, která jsou méně náročná na vytvoření i údržbu. Optimální by bylo dále zkoumat možnosti skloubení mapování takových hrubých schémat s automatickou indexací a také do jaké míry a jakým způsobem takto přidělené heslo PSH ovlivní výsledky případné automatické indexace. Využití mapování tezauru MeSH nebo obdobně podrobných schémat by sice významně obohatilo záznamy v systému NUŠL, bylo by ale dobré otestovat, zda vyhledávání v takto upravených záznamech přináší kvalitnější výsledky, či nikoli.

V případě analýzy možného využití obou metod sjednocování věcného popisu na autorských volných klíčových slovech popisujících vysokoškolské kvalifikační práce se mapování jevilo jako velice neefektivní řešení (více než 96 % klíčových slov není použito ani 10krát). Pro sjednocení volných klíčových slov se tedy v současné chvíli jeví automatická indexace jako vhodnější a jediná reálně proveditelná a udržitelná varianta.

Použitá literatura

About BASE: Statistics. UNIVERSITÄTSBIBLIOTHEK BIELEFELD. *BASE* [online]. 2016 [cit. 2016-05-12]. Dostupné z: http://www.base-search.net/about/en/about_statistics.php?menu=2

About OpenGrey. *OpenGrey* [online]. 2011 [cit. 2015-09-17]. Dostupné z: http://opengrey.eu/about

BALÍKOVÁ, Marie, 2003a. Selekční jazyk. In*: KTD: Česká terminologická databáze knihovnictví a informační vědy (TDKIV)* [online]. Praha: Národní knihovna ČR [cit. 2016-05-31]. Dostupné z: http://aleph.nkp.cz/F/?func=direct&doc_number=000001625&local_base=KTD.

BALÍKOVÁ, Marie, 2003b. Dokumentační selekční jazyk. In: *KTD: Česká terminologická databáze knihovnictví a informační vědy (TDKIV)* [online]. Praha: Národní knihovna ČR [cit. 2016-05-31]. Dostupné z: http://aleph.nkp.cz/F/?func=direct&doc_number=000001519&local_base=KTD.

BALÍKOVÁ, Marie, 2003c. Věcný selekční jazyk. In: *KTD: Česká terminologická databáze knihovnictví a informační vědy (TDKIV)* [online]. Praha: Národní knihovna ČR [cit. 2016-05-31]. Dostupné z: http://aleph.nkp.cz/F/?func=direct&doc_number=000001659&local_base=KTD.

BALÍKOVÁ, Marie, 2003d. Hierarchický vztah. In: *KTD: Česká terminologická databáze knihovnictví a informační vědy (TDKIV)* [online]. Praha: Národní knihovna ČR [cit. 2016-05-31]. Dostupné z: http://aleph.nkp.cz/F/?func=direct&doc_number=000001540&local_base=KTD.

BALÍKOVÁ, Marie, 2003e. Nadřazený termín. In: *KTD: Česká terminologická databáze knihovnictví a informační vědy (TDKIV)* [online]. Praha: Národní knihovna ČR [cit. 2016-06-01]. Dostupné z: http://aleph.nkp.cz/F/?func=direct&doc_number=000001579&local_base=KTD.

BALÍKOVÁ, Marie, 2003f. Podřazený termín. In: *KTD: Česká terminologická databáze knihovnictví a informační vědy (TDKIV)* [online]. Praha: Národní knihovna ČR [cit. 2016-06-01]. Dostupné z: http://aleph.nkp.cz/F/?func=direct&doc_number=000001594&local_base=KTD.

BALÍKOVÁ, Marie, 2003g. Předmětová kategorizace informačních zdrojů pro potřeby Konspektu. In: *Národní knihovna.* **14**(1), s. 42-54. ISSN 0862-7487. Dostupné také z: http://oldknihovna.nkp.cz/NKKR0301/0301042.html.

BALÍKOVÁ, Marie, 2003h. Deskriptorový odstavec. In: *KTD: Česká terminologická databáze knihovnictví a informační vědy (TDKIV)* [online]. Praha: Národní knihovna ČR [cit. 2016-06-01]. Dostupné z: http://aleph.nkp.cz/F/?func=direct&doc_number=000001517&local_base=KTD.

BALÍKOVÁ, Marie, 2003i. Asociativní vztah. In: KTD: Česká terminologická databáze knihovnictví a informační vědy (TDKIV) [online]. Praha: Národní knihovna ČR [cit. 2016-07-06]. Dostupné z: http://aleph.nkp.cz/F/?func=direct&doc_number=000001504&local_base=KTD.

BALÍKOVÁ, Marie, 2012. Varianty věcného zpřístupnění informačních zdrojů. *In: Národní knihovna ČR* [online]. [cit. 2016-06-01]. Dostupné z: http://text.nkp.cz/o-knihovne/odborne-cinnosti/zpracovani-fondu/roztridit/zazn-zpristinfzdroj

BAWDEN, David a ROBINSON. *An introduction to information science*. London: Facet, 2012. ISBN 18-560-4810-1.

BRATKOVÁ, Eva a Helena KUČEROVÁ, 2014. Systémy organizace znalostí a jejich typologie. *Knihovna: knihovnická revue.* **25**(2), s. 5-29. ISSN 1801-3252. Dostupné také z: http://knihovna.nkp.cz/knihovna142/142095.htm.

EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH, 2008. *BibClassify Admin Guide* [online]. [cit. 2016-05-31]. Dostupné z: https://cds.cern.ch/help/admin/bibclassify-admin-guide

FARACE, Dominic John a Joachim SCHÖPFEL, 2010. *Grey literature in library and information studies*. New York: De Gruyter Saur.

FRANTÍKOVÁ, Bohdana, 2014. *Pokyny pro zpracování záznamů v systému Invenio* [online]. Praha: NTK [cit. 2016-05-04]. Dostupné z: http://invenio.nusl.cz/record/111520?ln=cs. Dokument je průběžné aktualizován. Citována byla verze z roku 2014.

HENDL, Jan. *Přehled statistických metod: analýza a metaanalýza dat*. Páté, rozšířené vydání. Praha: Portál, 2015. ISBN 978-80-262-0981-2.

HRAZDIL, Aleš, 2003. Paradigmatický vztah. In: *KTD: Česká terminologická databáze knihovnictví a informační vědy (TDKIV)* [online]. Praha: Národní knihovna ČR [cit. 2016-05-31]. Dostupné z: http://aleph.nkp.cz/F/?func=direct&doc_number=000001587&local_base=KTD.

INCAD, 2013. *Automatická indexace PSH PSHASSIGN: manuál*. Praha. Interní dokument.

ISO 25964-2:2013. Information and documentation – Thesauri and interoperability with other vocabularies – Part 2: Interoperability with other vocabularies. 1st ed. Geneva: International Organization for Standardization, 2013-03-04. 99 s.

JANSKÁ, Lenka, 2008. Vícejazyčné tezaury. In: *Inflow* [online]. [cit. 2016-05-31]. Dostupné z: http://www.inflow.cz/vicejazycne-tezaury#_ftn4.

JIHOČESKÁ UNIVERZITA, 2016. O Jihočeské univerzitě. In: *Jihočeská univerzita* [online]. [cit. 2016-06-01]. Dostupné z: https://www.jcu.cz/o-univerzite.

KOCOUREK, Pavel, 2013. Automatická indexace šedé literatury hesly Polytematického strukturovaného hesláře. In: *Seminář ke zpřístupňování šedé literatury 2013 : 6. ročník semináře zaměřeného na problematiku uchovávání a zpřístupňování šedé literatury, 23. 10. 2013* [online]. Praha: Národní technická knihovna, [cit. 2016-05-31]. Dostupný z: https://invenio.nusl.cz/record/161469.

KOŽUCHOVÁ, Kristýna a Ctibor ŠKUTA, 2010. Polytematický strukturovaný heslář a jeho potenciál v oblasti třídění a zpřístupňování webových dokumentů. In: *INFORUM 2010: 16. ročník konference o profesionálních informačních zdrojích Praha, 25.–27. 5. 2010* [online]. Praha [cit. 2016-05-04]. Dostupné z: http://www.inforum.cz/sbornik/2010/67/. Dostupné také z: http://repozitar.techlib.cz/record/33/.

KOŽUCHOVÁ, Kristýna, 2012. *Zásady pro vytváření Polytematického strukturovaného hesláře (PSH)* [online]. Praha: Národní technická knihovna [cit. 2016-05-31]. Dostupné z: http://repozitar.techlib.cz/record/728.

KUČEROVÁ, H., 2005. *Současný stav a postavení Polytematického strukturovaného hesláře (PSH) mezi ostatními selekčními jazyky: srovnávací analýza*. Praha. Interní materiál.

LICHTENBERGOVÁ, E, 2000. Dotazy ke katalogizaci. In: *Národní knihovna ČR* [online]. [cit. 2016-06-01]. Dostupné z: http://katdotaz.nkp.cz/zobraz.phtml?id=319.

LÖSCH, M, 2009. *Automatische Klassifikation von OAI-Metadaten mit linguistischen Methoden* [online]. [cit. 2015-09-17]. Dostupné z: http://129.70.12.22//wikifarm/fields/ub_edv/uploads/Oeffentlich/auto_oai_slides.pdf

LÖSCH, M, 2011. *Automatische Sacherschließung elektronischer Dokumente*. [online]. [cit. 2015-09-17]. Dostupné z: https://opus4.kobv.de/opus4-bib-info/frontdoor/index/index/docId/1001

LÖSCH, M., U. WALTINGER, W. HORSTMANN a A. MEHLER, 2011. Building a DDC-annotated Corpus from OAI Metadata. *Journal of Digital Information*. **12**(2) [cit. 2015-09-17]. Dostupné z: https://journals.tdl.org/jodi/index.php/jodi/article/view/1765.

MAIXNEROVÁ, Lenka a Alena SMUTNÁ, 2014. Aktualizace tezauru MeSH - verze 2015. In: *Lékařská knihovna: časopis pro knihovny a informační střediska ve zdravotnictví* [online], **19**(3-4) [cit. 2016-06-01]. ISSN 1804-2031. Dostupné z: http://www.nlk.cz/publikace-nlk/lekarska-knihovna/2014/19-3-4/aktualizace-tezauru-mesh.

MERLIN Project Proposal: MERLIN Metadata Enrichment for Repositories in a London Institutional Network [online]. 2008 [cit. 2015-09-17]. Dostupné z: https://code.google.com/p/jisc-merlin/downloads/detail?name=MERLIN_proposal_public.pdf&can=2&q=

MOYLE, Martin, 2011. *MERLIN Final project report* [online]. [cit. 2015-09-17]. Dostupné z: http://discovery.ucl.ac.uk/1321559/

MOYLE, M., R. STOCKLEY a S. TONKIN. SHERPA-LEAP: a consortial model for the creation and support of academic institutional repositories. *OCLC systems and services: international digital library perspective ;including a special section : institutional repositories* [online]. Bradford, England: Emerald Group, 2007, **23**(2): 125-132 [cit. 2015-09-17]. Dostupné z: http://discovery.ucl.ac.uk/2663/

MYNARZ, Jindřich, 2009. Jak lze prakticky využít Polytematický strukturovaný heslář pro věcný popis elektronických zdrojů. *Ikaros*[online]. **13**(12) [cit. 2016-05-31]. urn:nbn:cz:ik-13285. ISSN 1212-5075. Dostupné z: http://ikaros.cz/node/13285.

MYNARZ, Jindřich a Ctibor ŠKUTA, 2010. Integration of an Automatic Indexing System within the Document Flow of a Grey Literature Repository. In: FARACE, D. J. a J. FRANTZEN, compil. *Transparency in Grey Literature, Grey Tech Approaches to High Tech Issues: Twelfth International Conference on Grey Literature, 6-7 December 2010 in the National Technical Library, Prague, Czech Republic: GL 12 conference proceedings.* Amsterdam: TextRelease, February 2011. 142 s. GL-Conference series, ISSN 1386-2316, no. 12. ISBN 978-90-77484-16-6. ISBN 90-77484-16-7. Dostupné také z: http://invenio.nusl.cz/record/42005?ln=cs

MYNARZ, Jindřich, Ctibor ŠKUTA a Tomáš MÜLLER, 2011. *Jak dokumentům automaticky přiřadit hesla PSH* [online]. Praha [cit. 2016-05-31]. Dostupné z: http://repozitar.techlib.cz/record/105. Prezentace z konference Co se skrývá za vyhledáváním aneb Searching Session NTK 2011, Praha (CZ), 2011-10-04.

NÁRODNÍ LÉKAŘSKÁ KNIHOVNA, 2015. Struktura MeSH. In: *Národní lékařská knihovna* [online]. [cit. 2016-06-01]. Dostupné z: http://www.nlk.cz/informace-o-nlk/odborne-cinnosti/tezaurus-medical-subject-headings/struktura-mesh.

NÁRODNÍ TECHNICKÁ KNIHOVNA, 2012. Automatická indexace obsahu Národního úložiště šedé literatury pomocí Polytematického strukturovaného hesláře. In: *Gemin: elektronické tržiště* [online]. [cit. 2016-05-31]. Dostupné z: https://www.gemin.cz/index.php?id=3303&state=PREP&step=PREP&m=contracts&h=contract&a=dashboard#step_href_PREP_PREP.

NÁRODNÍ TECHNICKÁ KNIHOVNA, 2015. Polytematický strukturovaný heslář. NTK. *Národní technická knihovna* [online]. Praha, ©2006-2016 [cit. 2016-05-31]. Dostupné z: https://www.techlib.cz/cs/82897-polytematicky-strukturovany-heslar.

PALA, Karel, 1996. Informační technologie a korpusová lingvistika (2). *Zpravodaj ÚVT MU* [online]. **6**(4) [cit. 2015-09-17]. ISSN 1212-0901. Dostupné z: http://webserver.ics.muni.cz/bulletin/articles/67.html

PEJŠOVÁ, Petra, 2008. Projekt NUŠL a další projekty v ČR: Projekt „Digitální knihovna pro šedou literaturu – funkční model a pilotní realizace“. In: *Seminář ke zpřístupňování šedé literatury 2008: 1. ročník semináře zaměřeného na problematiku uchovávání a zpřístupňování šedé literatury, 8. 10. 2008* [online]. Praha: Národní technická knihovna [cit. 2016-05-04]. ISSN 1803-6015. Dostupný z: http://nusl.techlib.cz/konference/sbornik-2008/.

PEJŠOVÁ, Petra a Iveta FÜRSTOVÁ, 2010. Národní úložiště šedé literatury (NUŠL). In: *INFORUM 2010: 16. konference o profesionálních informačních zdrojích Praha, 25. - 27. 5. 2010* [online]. Praha [cit. 2016-05-04]. ISSN 1801-2213. Dostupné z: http://www.inforum.cz/sbornik/2010/70/. Dostupné také z: http://repozitar.techlib.cz/record/22/.

PÍŠKOVÁ, Milada. *Věcná katalogizace [online]*. Praha [cit. 2016-05-31]. Dostupné z: http://repozitar.techlib.cz/record/558

PRESOVÁ, Silvie, 2005. *Metoda Konspektu a její vztah k selekčním jazykům: současný stav a trendy se zaměřením na situaci v České republice*. 1. vyd. V Brně: Masarykova univerzita. 152 s. ISBN 80-210-3685-0.

SCHÖPFEL, Joachim, Christiane STOCK a Nathalie HENROT, 2007. *From SIGLE to OpenSIGLE and beyond: An in-depth look at Resource Migration in European Context* [online]. [cit. 2015-09-17]. Dostupné z: http://www.greynet.org/images/GL8,_page_47.pdf

SCHÖPFEL, Joachim, 2010. Towards a Prague Definition of Grey Literature. In: FARACE, D. J. a J. FRANTZEN, compil. *Transparency in Grey Literature, Grey Tech Approaches to High Tech Issues: Twelfth International Conference on Grey Literature, 6-7 December 2010 in the National Technical Library, Prague, Czech Republic: GL 12 conference proceedings.* Amsterdam: TextRelease, February 2011. 142 s. GL-Conference series, ISSN 1386-2316, no. 12. ISBN 978-90-77484-16-6. ISBN 90-77484-16-7. Dostupné také z: http://archivesic.ccsd.cnrs.fr/sic_00581570/document.

SKOLKOVÁ, Linda, 2007a. *Polytematický strukturovaný heslář*. Praha. Diplomová práce (Mgr.). Univerzita Karlova v Praze, Filozofická fakulta, Ústav informačních studií a knihovnictví. Dostupné také z: https://is.cuni.cz/webapps/zzp/detail/28014.

SKOLKOVÁ, Linda, 2007b. *Indexační pravidla pro práci s Polytematickým strukturovaným heslářem (PSH)* [online]. Praha, 2007 [cit. 2016-05-31]. Dostupné z: http://repozitar.techlib.cz/record/724.

SMOLKA, Pavel, 1998. PSH - polytematický strukturovaný heslář. *Národní knihovna*. **9**(3), s. 130-135. ISSN 0862-7487.

STOCK, Christiane a Nathalie HENROT, 2011. From OpenSIGLE to OpenGrey: changes and continuity. In: *Grey Journal*. **7**(2), 93-97. ISSN 1574-1796. Dostupné komerčně z EBSCO (LISS): https://search.ebscohost.com/. Dostupné také z: http://www.greynet.org/images/GL12_S3P,_Stock_and_Henrot.pdf

SUMMANN, Friedrich a Norbert LOSSAU, 2004. Search Engine Technology and Digital Libraries. *D-Lib Magazine* [online]. **10**(9) [cit. 2015-09-17]. DOI: 10.1045/september2004-lossau. ISSN 1082-9873. Dostupné z: http://www.dlib.org/dlib/september04/lossau/09lossau.html

ŠÍMOVÁ, Alena a Lenka MAIXNEROVÁ, 2008. Aktualizace českého překladu MeSH. *Ikaros* [online]. **12**(5) [cit. 2016-06-01]. urn:nbn:cz:ik-12778. ISSN 1212-5075. Dostupné z: http://ikaros.cz/node/12778.

UNIVERSITÄTSBIBLIOTHEK BIELEFELD, 2004a. FAQ. In: UNIVERSITÄTSBIBLIOTHEK BIELEFELD. *BASE* [online]. [cit. 2015-09-17]. Dostupné z: http://www.base-search.net/about/en/faq.php?#requirements.

UNIVERSITÄTSBIBLIOTHEK BIELEFELD, 2004b. Suggest repository / e-journal. In: UNIVERSITÄTSBIBLIOTHEK BIELEFELD. *BASE* [online]. [cit. 2015-09-17]. Dostupné z: http://www.base-search.net/about/en/suggest.php.

UNIVERSITÄTSBIBLIOTHEK BIELEFELD, 2004c. Suchmaschine BASE. In: UNIVERSITÄTSBIBLIOTHEK BIELEFELD. *BASE* [online]. [cit. 2015-09-17]. Dostupné z: http://www.base-search.net/

UNIVERSITÄTSBIBLIOTHEK BIELEFELD, 2010. Projektarchiv. In: UNIVERSITÄTSBIBLIOTHEK BIELEFELD, *UB Wiki* [online]. [cit. 2015-09-17]. Dostupné z: http://www.ub.uni-bielefeld.de/biblio/projects/oai_projekt.htm

UNIVERSITÄTSBIBLIOTHEK BIELEFELD, 2011. Projektergebnisse. In: UNIVERSITÄTSBIBLIOTHEK BIELEFELD, *UB Wiki* [online]. [cit. 2015-09-17]. Dostupné z: http://www.ub.uni-bielefeld.de/wiki/OAIMErgebnisse

UNIVERZITA PARDUBICE, 2015. Vyhledávání zaměstnanců. In: *Univerzita Pardubice* [online]. Pardubice [cit. 2016-05-31]. Dostupné z: http://www.upce.cz/lide-hledat.html.

WORLD WIDE WEB CONSORTIUM, 2009. *SKOS Simple Knowledge Organization System Reference* [online]. [cit. 2016-05-31]. Dostupné z: https://www.w3.org/TR/skos-reference/.

WALTINGER, U., LÖSCH, M., HORSTMANN, W. a A. MEHLER, 2010. Enhancement of OAI Metadata via Automatic Document Classification. [online]. Dostupné z: http://129.70.12.22//wikifarm/fields/ub_edvmain/uploads/Oeffentlich/loesch_gfkl_presentation.pdf

WALTINGER, Ulli, Alexander MEHLER, Mathias LÖSCH a Wolfram HORSTMANN, 2009. Hierarchical Classification of OAI Metadata Using the DDC Taxonomy. In: BERNARDI, Raffaella, SEGOND, Frederique a Ilya ZAIHRAYEU, ed. *Advanced language technologies for digital libraries: international workshops on NLP4DL 2009, Viareggio, Italy, June 15, 2009 and AT4DL 2009, Trento, Italy, September 8, 2009*. New York: Springer, s. 29-40. ISBN 978-3-642-23160-5.

Seznam vyobrazení a příloh

Obrázky

Tabulky

Seznam příloh na CD

Příloha 1: Tabulka s údaji ze záznamů vzorku ve formátu csv.

Příloha 2: Popis sloupců tabulky v příloze 1.