



VYSOKÁ ŠKOLA CHEMICKO-TECHNOLOGICKÁ V PRAZE
Fakulta chemické technologie

Laboratoř informatiky a chemie

Diplomová práce

**Automatizace procesů správy a možnosti integrace
Polytematického strukturovaného hesláře v Národní
technické knihovně**

**Automatization of management processes and
possibilities of integration of Polythematic Structured
Subject Heading System at National Technical Library**

Bc. Ctibor Škuta

Prohlašuji, že jsem předloženou diplomovou práci vypracoval samostatně a použil jsem jen pramenů, které cituji a uvádím v seznamu použité literatury.

V Praze dne

Podpis

Souhrn

Automatizace procesů správy a možnosti integrace Polytematického strukturovaného hesláře v Národní technické knihovně

Polytematický strukturovaný heslář (PSH) je česko-anglický řízený slovník, spravovaný Národní technickou knihovnou (NTK). Jeho primárním účelem je ruční indexace dokumentů v NTK a dalších paměťových institucích, které jej používají. Jeho současná verze (PSH 3.0) obsahuje přes 13 800 hesel, rozdělených do 44 kategorií pokrývajících hlavní okruhy lidského vědění. Cílem práce je automatizace a standardizace procesů spojených s jeho obsahovou aktualizací, správa a vývoj prostředků určených pro jeho prezentaci a vyhledávání možností integrace do projektů, které mohou těžit z jeho strukturních vlastností.

Automatization of management processes and possibilities of integration of Polythematic Structured Subject Heading System at National Technical Library

Polythematic structured subject heading system (PSH) is czech-english controlled vocabulary managed at the National technical library (NTL). PSH is primarily intended as a tool for manual indexing of documents at NTL and other memory institutions. Its current version (PSH 3.0) contains over 13 800 subject headings divided into 44 categories which cover the main aspects of human knowledge. The aim of this work is the automation and standardization of its content update and management including development of software tools. Integration of PSH to other projects is also investigated.

Studijní obor: **Aplikovaná informatika v chemii**

Diplomová práce: **Bc. Ctibor Škuta**

Vedoucí práce: **Ing. Miloslav Nič, Ph.D.**

Práce odevzdána dne:

Poděkování

Tímto bych chtěl poděkovat všem svým kolegům, kteří mi pomáhali s realizací úkolů, ať už prakticky, či jen teoreticky.

Jmenovitě Kristýně Kožuchové za úzkou spolupráci při správě hesláře a celkové zasvěcení do problematiky knihoven a systémů organizace znalostí. Dále Pavlíně Omastové za spolupráci při vyhodnocování výsledků při aktualizacím procesu a za její výsledné statistiky. Jindřichu Mynarzovi za pomoc při realizaci některých řešení i za praktickou spolupráci. Tomáši Müllerovi za pomoc při integraci automatického indexeru do digitálního repozitáře, který v NTK osobně spravuje.

Speciální poděkování patří mé vedoucí Radce Římanové, a to za lidskou podporu, přístup a vstřícnost při řešení pracovních záležitostí.

Samozřejmě nesmím zapomenout ani na svého školitele Miloslava Niče, který mi pomohl s celkovým uchopením diplomové práce a podporoval mě pohodovým přístupem v průběhu její tvorby.

Obsah

1	Úvod	1
1.1	Organizace znalostí	1
1.2	Předmětová indexace	2
2	Systémy organizace znalostí.....	3
2.1	Druhy systémů organizace znalostí.....	5
2.2	Správa znalostního systému.....	8
2.2.1	Aktualizace	8
2.2.2	Uživatelé jako zdroj aktualizací.....	8
2.2.3	Opravy	9
2.2.4	Pravidla pro tvorbu hesel	9
3	Polytematický strukturovaný heslář	12
3.1	Historie	13
3.2	Současnost.....	13
3.3	Využití PSH v institucích	14
3.4	Formáty PSH.....	15
3.5	MARC	15
3.5.1	Historie.....	15
3.5.2	MARC21 pro authority.....	16
3.5.3	Struktura formátu.....	16
3.5.4	MARC v současnosti	18
3.6	SKOS.....	19
3.6.1	Struktura formátu.....	20
3.6.2	Mapování hesel na vnější systémy	22
3.6.3	Výhody SKOS.....	22
3.6.4	Aplikace SKOS.....	23

3.7	Prohlížení hesláře	24
3.7.1	Vytváření úryvků metadat s hesly PSH	27
4	Cíl práce	29
5	Užité programovací technologie	31
5.1	Python	31
5.2	Django Framework.....	31
5.3	JavaScript	32
5.4	jQuery	32
5.5	MySQL	32
5.6	HTML	33
5.7	CSS	33
6	Procesy správy hesláře.....	34
6.1	Aktualizace PSH.....	34
6.2	Zdroje aktualizací	34
6.3	Formulář pro návrh nového hesla	34
6.4	Analýza klíčových slov	35
6.4.1	Výsledky analýzy klíčových slov	38
6.4.2	Následky analýzy klíčových slov	38
6.4.3	Sikuli.....	39
6.4.4	Programování v Sikuli	39
6.4.5	Výhody vizuálního přístupu.....	40
6.4.6	Nevýhody vizuálního přístupu	40
6.4.7	Automatizace Aleph GUI.....	41
6.4.8	Použitá řešení	41
6.4.9	Výsledky automatizace	42
6.5	Analýza LOGu vyhledávání v katalogu NTK	43

6.5.1	Zdroj informací	43
6.5.2	Analýza logů	43
6.5.3	Log knihovnického informačního systému Aleph	44
6.5.4	Technologický popis	45
6.5.5	Postup práce	47
6.5.6	Algoritmus analýzy	48
6.5.7	Výsledky analýzy	50
6.5.8	Současnost	50
7	Prezentace PSH	52
7.1	Budoucnost Prohlížení PSH	52
7.2	Nová prezentace PSH na stránkách NTK	53
7.2.1	Technologický popis	53
7.2.2	Současný stav	54
7.3	PSH Manager	55
7.3.1	Technologický popis	55
7.3.2	Uživatelské rozhraní	56
8	Nové možnosti využití hesláře	63
8.1	Automatická indexace dokumentů v NUŠL	63
8.1.1	Automatická indexace	64
8.1.2	Technologický popis	64
8.1.3	Proces indexace	67
8.1.4	Uživatelské rozhraní	68
8.1.5	Výsledky testování	68
8.1.6	Budoucí možnosti	69
9	Závěr	72
10	Seznam použité literatury	74

11 Přílohy..... 84

DVD příloha: Na přiloženém DVD naleznete PDF verzi bakalářské práce, včetně všech programových materiálů, tak jak jsou prezentovány v textu.

1 Úvod

Lidé se od dávných dob snaží organizovat a třídit věci kolem sebe. Dávají jim jména, klasifikují je – dalo by se říct, že v nich dělají pořádek. Tato činnost často souvisí se shlukováním podobných objektů do vyšších skupin - ukládáním na stejná místa. Na rozdíl od skutečných věcí, které se v danou chvíli mohou nacházet fyzicky jen na jednom místě, teoretické přiřazování popisných termínů nám umožňuje zařadit konkrétní věc do nekonečného množství množin objektů, a to podle různých hledisek.

1.1 Organizace znalostí

S narůstajícím povědomím o přírodních zákonech se původně jediná vědní disciplína, filosofie, začala dále diferencovat na specializovaná odvětví lidského vědění podle hlavního aspektu, jakým nahlížíme na náš svět. Tehdy začalo docházet k organizaci abstraktního světa znalostí, která pokračuje dodnes a pokračovat bude. Některé z vědních oborů se budou dále prohlubovat a budou vznikat další, nové.

Knihovny jsou odjakživa považovány za jedna z hlavních center lidského vědění. Asi nejznámějším příkladem je slavná alexandrijská knihovna, která se snažila soustředit skutečně veškeré lidské vědění své doby. To se jí, jak známo, dařilo až do roku 47 př. n. l., kdy částečně vyhořela, což bylo následováno nešťastnými událostmi staletí nadcházejících.^{a 1} Před dvěma tisíci lety navíc bylo vzhledem k míře poznání přírodních zákonitostí možné mít prakticky vše na jednom místě.

V dnešní době to však možné není. Někdo by mohl namítnout, že takové místo je a je jím internet. Internet je však decentralizovaný a jediné, co bychom mohli říci je, že je zde obrovské množství důležitých informací, avšak mezi několikanásobně větším množstvím informačního odpadu. Informace na internetu navíc nejsou

^a Alexandrijská knihovna byla největší a nejslavnější knihovnou starověku. V době svého největšího rozkvětu obsahovala přes 700 000 rukopisů. V roce 47 př. n. l. částečně vyhořela za války mezi Caesarem a Pompeiem. V roce 389 byla zcela zbořena římskými vojsky. Po znovuvybudování pokračovala ve své činnosti až do roku 613, kdy byla Alexandrie dobyta araby.

žádným způsobem tříděny, uspořádány či kategorizovány. V tom nám pomáhají internetové vyhledávače, které filtrují obsah na základě našich požadavků (přesto s nejistým výsledkem) nebo informační portály, které samy informace třídí, poskytují nebo hodnotí jejich relevanci.

Třídění a selekce relevantních (rozumějme kvalitních) informací by měla být jednou z esenciálních funkcí knihoven, a také je. Jejich přínosem pro lidskou společnost tedy už velmi často není poskytnutí přístupu ke znalostem, které bychom nemohli získat jinou, kratší nebo pohodlnější cestou. Jejich hlavním přínosem je, že poskytují přístup k informacím kvalitním a kvalitně zpracovaným, protože to je práce, kterou knihovníci dělají odjakživa.²

1.2 Předmětová indexace

K třídění dokumentů dochází na základě přiřazení deskriptivních termínů (*indexů, deskriptorů*), tomuto procesu se tedy říká indexace. Indexem je každý popisný prvek dokumentu - název, autor či žánr. To jsou však údaje, které jsou dané a samozřejmé. Přidanou hodnotou je v tomto případě indexace předmětová (*klasifikace, kategorizace*). Předmětová indexace přiřazuje dokumentu jeden nebo více termínů, které by co nejlépe měly vystihovat jeho věcný obsah.

V ideálním případě jsou podobné dokumenty popisovány stejnými hesly, které zajistí, že výsledek vyhledávání v katalogu je konzistentní, ale také obsáhlý. Zde bychom však mohli narazit na problém lidského faktoru. Je více než logické, že ne všichni indexátoři danému dokumentu přiřadí stejné deskriptivní termíny a pravděpodobně ani člověk samotný nezvolí stejná hesla v případě, že dokument indexuje opakovaně po delší době. Nehledě na to, že ke každému termínu najdeme několik termínů synonymických a že ke klasifikaci může docházet na různé úrovni předmětového popisu (např. obecný popis – *chemie*, přesnější popis – *organická chemie*). Za účelem minimalizace těchto nechtěných jevů se v knihovnách i všude jinde, kde dochází ke klasifikaci, používají různé druhy systémů organizace znalostí.

2 Systémy organizace znalostí

“Termín systém organizace znalostí zahrnuje všechny typy schémat určených pro třídění informací a posilování znalostního managementu. Systémy organizace znalostí zahrnují klasifikační a kategorizační schémata, která organizují materiály na základní úrovni, předmětová hesla, která poskytují detailnější popis, a autoritní soubory, které spravují různé druhy klíčových informací, jako jsou geografické názvy nebo osobní jména. Systémy organizace znalostí zahrnují také vysoce strukturované slovníky, jako jsou tezaury, a méně tradiční schémata, kterými jsou sémantické sítě a ontologie. Systémy organizace znalostí jsou mechanismy pro třídění informací, a proto jsou srdcem každé knihovny, muzea a archivu.“ [HODGE, 2000]³

Systémem organizace znalostí (*Knowledge organization system*, KOS) rozumíme omezenou množinu termínů, na jejímž základě rozdělujeme určitá data do kategorií. Nejsou typické pouze pro knihovny a paměťové instituce, ale pro všechny, co nějakým způsobem třídí a shlukují svá data. Můžeme je vidět všude kolem sebe a nemusí se nutně jednat o třídění knih nebo textů. Jedním z příkladů může být rozdělení výrobků do kategorií v internetovém obchodě, dalším zařazení dřeviny mezi stromy a dále mezi listnaté stromy.

Jak je vidět, termíny tvořící systém organizace znalostí mohou existovat samostatně - bez vazby na okolí, nebo mohou dohromady tvořit provázanou strukturu, nejčastěji strom. Základními vazbami jsou vazby hierarchické – vztah nadřazené heslo, podřazené heslo, vazby ekvivalentní – hesla synonymická, a vazby příbuzné – hesla spolu souvisí, ale ve struktuře se nachází na různých místech.

Termíny tvořící provázanou strukturu pro nás samozřejmě mají mnohem větší hodnotu než samostatná hesla. Určují totiž nejen obsah dokumentu, ale zároveň určují jeho vztah k dokumentům ostatním. Na jejich základě tak můžeme hledat nejen texty podobné, ale také obecnější nebo detailnější.

To ovšem neznamená, že bychom na základě provázanosti či počtu termínů mohli říct, že je jeden systém lepší než druhý. Důležitý je vždy kontext a účel využití.⁴

Z těchto důvodů se v knihovnách často setkáváme s použitím více klasifikačních systémů najednou. Jeden z nich může určovat fyzické umístění dokumentu ve fondu knihovny (pozici určuje jediné heslo), zatímco druhý lépe poslouží při indexaci pro budoucí vyhledání v katalogu (dokument popsán zpravidla více hesly).⁵

Dalším aspektem při rozlišování klasifikačních systémů je jejich tematické zaměření. Zaměření by se opět mělo odvíjet od účelu použití, v našem případě nejčastěji fondu knihovny, který navíc určuje směr vývoje daného systému. V technické knihovně se logicky budou více rozvíjet a aktualizovat technicky zaměřené části systému, zatímco v knihovně biologického ústavu budou více rozvíjeny větve přírodovědné.

I přes velkou diverzitu však sdílí všechny systémy organizace znalostí základní charakteristiku:

- Systém organizace znalostí představuje určitý pohled na danou kolekci a její data.
- Stejně entity mohou být charakterizovány z různých pohledů, které závisí na použitém systému.
- Musí existovat dostatečná shoda mezi konceptem vyjádřeným v systému a skutečným objektem, na který tento koncept odkazuje, tak aby jej člověk mohl rozumně a spolehlivě používat. Stejně tak, osoba používající systém pro vyhledávání musí být schopna spojit skutečný objekt s jeho reprezentací.

2.1 Druhy systémů organizace znalostí

2.1.1.1 Jednoduché systémy

Jak bylo již zmíněno, nejzákladnější druhy systémů organizace znalostí jsou systémy výčtové neboli množiny samostatných hesel bez vazby na sousední hesla.

Rejstřík (glosář) je ve většině případů abecedně seřazený soubor termínů z daného oboru, sloužící jako příloha specializovaného textu.

Slovník je typ systému, který organizuje informace o slovech. Může se jednat o jejich překlad, význam, původ, výslovnost aj.

Gazeteer je rejstříkem geografických dat (součást atlasů). K jednotlivým termínům ve mnoha případech poskytuje polohu a další doplňkové informace.

Autoritní soubor je seznam schválených standardizovaných hesel používaných při popisu dat při katalogizaci či inventarizaci. Obsahuje rovinu synonymních termínů, které odkazují na preferované, primární znění hesla.

2.1.1.2 Středně pokročilé systémy

Dalšími systémy jsou schémata hesel s omezenými druhy vztahů. Jedná se o nejčastější typy systémů organizace znalostí používaných v knihovnách po celém světě.

Předmětová hesla představují množinu kontrolovaných termínů, které popisují položky v konkrétní kolekci. Často pokrývají širokou škálu kategorií, ale jejich struktura bývá omezená. Mívají svá vlastní pravidla tvorby i používání. Nejznámějšími systémy předmětových hesel jsou Lékařská předmětová hesla (*Medical Subject Headings, MeSH*) a Předmětová hesla Kongresové knihovny^a (*Library of Congress Subject Headings, LCSH*).

Klasifikační schémata, taxonomie, kategorizační schémata zachycují strukturu klasifikačního stromu. Některé z nich poskytují číselné nebo abecední notace k

^a Kongresová knihovna je výzkumnou knihovnou Kongresu Spojených států amerických a de facto národní knihovnou Spojených států. Jedná se o nejstarší federální kulturní instituci ve Spojených státech. Její tři budovy ve Washingtonu D.C. představují největší knihovnu na světě do počtu regálů i počtu knih.

reprezentaci širokých oborů. Z knihovnického prostředí je nejznámější Klasifikace Kongresové knihovny (*Library of Congress Classification, LCC*) a Deweyho decimální klasifikace nebo také Deweyho decimální systém (*Dewey decimal classification, DDC*), který byl přímo vytvořen pro co nejjednodušší způsob uspořádání knih do regálů.

2.1.1.3 Pokročilé systémy

Jako poslední zmíníme nepokročilejší systémy organizace znalostí z hlediska struktury a komplexnosti vztahů mezi hesly/koncepty.

Tezaury jsou založeny na systému konceptů s vnitřním propojením. Obsahují tři základní druhy vztahů (ale může jich být i více nebo méně) – hierarchické, ekvivalenční a asociativní, každý z nich většinou uveden definovanou notací. Tezaury mají definovaná pravidla tvorby, která se však mohou individuálně lišit. Bývají velmi rozsáhlé, ale také specializované pro určitý obor.

Sémantická síť^a je systém hesel, na který můžeme nahlížet jako na orientovaný graf. Hesla jsou vzájemně propojena ohodnocenými vazbami, které vyjadřují logiku vztahu těchto objektů. Vztahy se tedy neorientují pouze na hierarchické nebo asociativní, ale bývají specifitější (můžou například určovat vtaž mezi dvěma lidmi v rámci rodiny nebo na základě osobních pocitů). Nejznámější typ sémantické sítě je WordNet^b. WordNet je lexikální databáze jazyka, která se skládá z množin kognitivních synonym, tzv. synsetů, které jsou vzájemně provázány. Popsány jsou i vztahy mezi samotnými slovy včetně definice slov a ukázkových vět.

Ontologie^c jsou nejmodernějšími systémy organizace znalostí vyjadřujícími formální logiku systému. Ontologie jsou popisem určité problematiky. Bývají spojovány především v souvislosti se sémantickým webem a umělou inteligencí. Ontologie obsahují čtyři základní prvky: jedince, třídy, atributy a vztahy. Jedinec je základním objektem ontologie. Může se jednat o člověka, zvíře i abstraktní věc jako

^a Sémantika (od řeckého séma – znak, znamení) je nauka o významu slov, znaků a jejich vztahu ke skutečnosti.

^b <http://wordnet.princeton.edu/>

^c Původní významem slova ontologie bylo označení filosofické disciplíny, která se zabývá bytím a základními pojmy (nejobecnějšími otázkami).

například číslo. Třída je množinou jedinců. Může obsahovat také podtřídy. Atributy jsou pak vlastnostmi jedinců nebo tříd. Vztahy v ontologiích mohou být jednosměrné i obousměrné, a nabývat velkého množství hodnot.

2.2 Správa znalostního systému

Správa znalostního systému (obr. 2.1) představuje komplexní soubor mechanismů, který je závislý na možnostech, prostředcích a zdrojích, které mohou správci využívat. Obecně by se však dal tento neustále probíhající proces rozdělit na dva hlavní úkoly, které spolu velmi úzce souvisí. Prvním z nich je aktualizace systému a druhým jeho opravy. K oběma procesům většinou dochází současně, často jeden z druhého vychází.

2.2.1 Aktualizace

Při prvotní tvorbě znalostního systému je patrně nejčastějším zdrojem hesel odborná literatura. Při jejich zařazování do organizační struktury se můžeme inspirovat v systému podobném (obsahově stejně zaměřeném), kde již konkrétní heslo existuje. Příliš velká inspirace však v tomto procesu není žádoucí. Systém netvoříme proto, aby se co nejvíce podobal jinému, a nezařazujeme nová hesla jen proto, že v jiném systému jsou a v našem ne (v takovém případě můžeme tento systém používat místo našeho). K zařazení nového hesla by mělo vždy dojít až po zralé úvaze s ohledem na účel a obsah, tedy kontext konkrétního systému.⁶

Systematické vyhledávání nových termínů je typické v raných fázích vývoje. U systémů v pokročilém stádiu však k tomuto procesu většinou nedochází. Jejich velikost to zpravidla neumožňuje, důsledkem by byla jen narůstající nekonzistence v různých částech.⁷ Odborné zdroje tak slouží především při zařazování termínů vycházejících z používání znalostního systému.

2.2.2 Uživatelé jako zdroj aktualizací

Hlavním zdrojem aktualizací používaného systému by vždy měli být jeho uživatelé. Uživatelé, kteří jej zprostředkovaně využívají za účelem vyhledat daný dokument, stejně tak jako aktivní uživatelé, kteří přiřazují hesla k indexovaným datům.⁸

Běžní uživatelé ani velmi často nevědí, že právě nalezená data jsou následkem jejich indexace za pomoci znalostního systému. Ale to není primárním účelem. Účelem je, že naleznou to, co hledají.

Uživatelé také sami nepřímým způsobem přispívají k rozvoji systému samotným vyhledáváním. Vyhledávání odráží potřeby, současné trendy, vývoj termínů v čase a postupy při

informačně-vyhledávacím procesu. Analýza chování uživatelů je jedním ze základních zdrojů aktualizace znalostních systémů.⁹

Hlavními iniciátory rozvoje systému po obsahové stránce, by měli být katalogizátoři. To hlavně z důvodů hluboké znalosti organizačního systému i indexovaných dat. Měli by navrhnout nová hesla, která v systému chybí a upozorňovat na konflikty v jeho struktuře. K této činnosti dochází také nepřímo v podobě klíčových slov.

Klíčová slova jsou termíny, které se ve schématu nevyskytují, ale indexátor je považuje za nejvhodnější deskriptory dokumentu. K tomu je tedy přiřadí, ale jako volně tvořený termín, ne standardizované předmětové heslo. Seznamy klíčových slov by měli procházet periodickou analýzou, která určí termíny vhodné k posouzení jako kandidáty na nová předmětová hesla.

2.2.3 Opravy

Opravy systémů ve většině případů vycházejí z problémů zjištěných při aktualizaci. Jejich aktivní vyhledávání je z hlediska schémat s velkým počtem termínů velmi neefektivní. Při opravách struktury je nutné být opatrný a brát v potaz všechny vazby vedoucí k danému konceptu. Opatrnost je na místě hlavně při odstraňování starších hesel. Tato změna se dotýká nejen schématu jako celku, ale i dat k němu navázaných. U systému s jedním aktivním správcem i uživatelem se tento problém minimalizuje. U systémů s centrální správou a více uživateli musíme brát ohled na data všech a o těchto změnách informovat nebo je i konzultovat.

2.2.4 Pravidla pro tvorbu hesel

Při přidávání hesel je nutné dbát na pravidla tvorby platící pro individuální systém. Všechna hesla by tedy měla mít stejnou jazykovou formu. Některé systémy obsahují pouze hesla v jednotném čísle, jiné upřednostňují číslo množné nebo se řídí podle toho, co zní přirozeně.

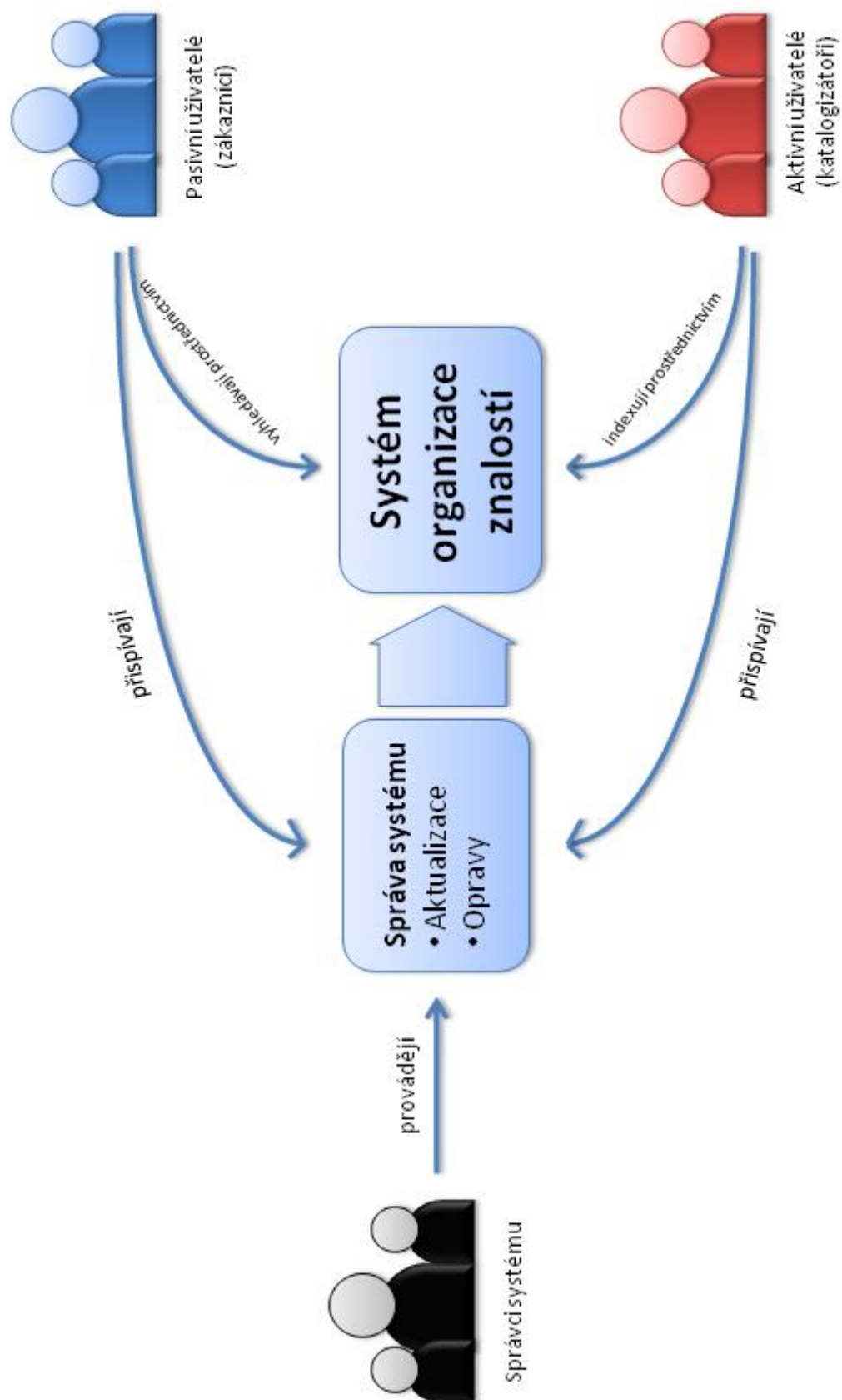
Pravidla se mohou lišit při manipulaci se slovy přenesenými. Zdomácnělé termíny jsou v řadě případů používanější než názvy národní (*smartphone* x *chytrý mobilní*

telefon). Řešení konfliktů je závislé na specifikacích systému – vícejazyčnost, přítomnost ekvivalentních termínů^a.

Důležitý je způsob, jakým se systém stará o homonyma. V ideálním případě je primární reprezentací hesla identifikátor. Slovní tvary se tak mohou pro různé kontexty shodovat. Dalším řešením je uvedení kontextu v textové reprezentaci, např. termínem v závorce, který jej upřesňuje. Poslední možností je úplné zamezení homonym, které však může nutit k vytváření deskriptorů v nepřirozeném tvaru.

Dalších pravidel může být celá řada od způsobu tvorby ekvivalentních termínů po aktualizaci zastaralých termínů. Hlavním pravidlem i pro všechna ostatní je co nejvyšší zachování konzistence systému.

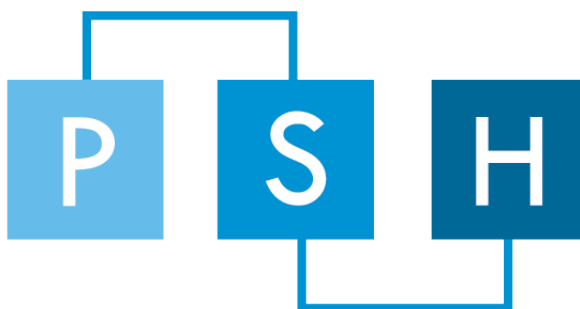
^a Ve vícejazyčných systémech i systémech s ekvivalentními hesly by měl být vždy hlavní národní termín.



Obr. 2.1 Diagram procesů ovlivňujících organizační systém

3 Polytematický strukturovaný heslář

Polytematický strukturovaný heslář (PSH) je česko-anglický řízený slovník lexikálních jednotek, který je primárně určen k předmětové indexaci dokumentů. V současné době je spravován referátem PSH v Národní technické knihovně (NTK). PSH je univerzální, což znamená, že se snaží pokrývat všechny okruhy lidského vědění. Je rozdělen do 44 tematických skupin, v nichž jsou hesla dále hierarchicky členěna v rámci nejvýše sedmistupňové hierarchie. Ve srovnání s klíčovými slovy či nekontrolovanými termíny, které jsou vytvářeny volně, má PSH pevnou strukturu i obsah. Hesla jsou tvořena na základě přirozeného jazyka a přirozeného pořádku slov, čímž umožňuje snadnější a efektivnější vyhledávání.



Obr. 3.1 Logo PSH

PSH patří mezi tzv. selekční jazyky, tj. mezi „umělé informační jazyky používané k vyjádření identifikačních nebo obsahových selekčních údajů za účelem organizace a vyhledávání dokumentů“ [TDKIV, selekční jazyk]. Konkrétně jej lze zařadit mezi předmětové selekční jazyky. Charakteristickým znakem této skupiny selekčních jazyků je postkoordinovanost. To znamená, že hesla jsou dokumentům přiřazována nezávisle na sobě a k jejich kombinaci dochází až při vyhledávání.¹⁰

3.1 Historie

Počátky práce na hesláři začaly v rozmezí let 1991 - 1993 ve Státní technické knihovně. Jeho vznik byl iniciován rozmachem automatizovaných knihovnických systémů.¹⁵ Vývoj byl několikrát podpořen granty Ministerstva kultury a Ministerstva školství, mládeže a tělovýchovy (MŠMT). Odborníci v jednotlivých oborech byli osloveni, aby zpracovali kategorické rozdělení hlavních témat, které jsou do současnosti základem nejvyšší vrstvy PSH. Tato práce trvala čtyři roky, aby v roce 1997 byla verze PSH 1.0 distribuována všem zájemcům mezi knihovnami i komerčními firmami.¹¹

V roce 2000 se heslář opět za podpory grantu od MŠMT dočkal svého překladu do angličtiny (PSH ver. 1.4). Tento krok měl přiblížit zapojení českých knihoven do mezinárodního prostředí a zároveň vylepšit možnosti přístupu ke knihovním sbírkám. Jazyková inovace byla logická z hlediska velkého množství cizojazyčných knih, zejména v odborné sféře, kde je angličtina nezbytným základem.¹² Nová dvojjazyčná verze hesláře byla následně rozhodnutím ministerstva zpoplatněna pro komerční sféru.

V roce 2005 byl uvolněn PSH ve verzi 2.0, která nebyla kvůli mnoha problémům distribuována. Důvodem bylo nedodržení základních pravidel pro aktualizaci hesláře při předchozích revizích (2002-2005). Odstraňovaná hesla nebyla po zrušení přiřazena jako nedeskriptory k existujícím preferovaným zněním, což způsobilo nekompatibilitu hesláře s bibliografickými záznamy. Problém byl zjištěný v roce 2006 po implementaci PSH do knihovního systému KP-Win, používaného ve Státní technické knihovně.¹³ Po revizi v roce 2006 byl PSH ver. 2.1 implementován do automatizovaného knihovnického systému Aleph. Ten slouží do dnešní doby jako primární software pro správu PSH.

3.2 Současnost

Heslář je v současné době spravován referátem PSH v Národní technické knihovně a jeho údržba probíhá v prostředí softwaru Aleph 500 ver. 18. Referát PSH byl v rámci organizačních změn souvisejících s transformací Státní technické knihovny na Národní technickou knihovnu zařazen do Odboru projektů a inovací.

Současná verze PSH 3.0 obsahuje přes 13 800 hesel preferovaných a přibližně 6000 hesel ekvivalentních. K aktualizaci dochází souvisle v průběhu roku, zatímco nové verze jsou uvolňovány na jeho začátku.

3.3 Využití PSH v institucích

Heslář mohou zcela zdarma využívat např. knihovny univerzitní, vědecké, odborné (při specializovaných ústavech a institucích, muzeích atd.), výjimečně i městské knihovny, knihovny občanských sdružení a jiné. Jednotlivec se stává uživatelem PSH zprostředkovaně, nejčastěji při vyhledávání v katalogu knihovny, která jej aktivně využívá.

Konkrétními aktivními uživateli PSH jsou kromě **Národní technické knihovny** např. **knihovny Českého vysokého učení technického v Praze^a**, **Ústřední knihovna Vysokého učení technického v Brně^b**, **Vědecká knihovna v Olomouci^c**, **knihovna Západočeského muzea v Plzni^d**, **Ústřední knihovna Filozoficko-přírodovědecké fakulty Slezské univerzity v Opavě^e**, **Knihovna Bankovního institutu vysoká škola v Praze^f** či **Univerzitní knihovna Pardubice^g**.

^a <http://knihovny.cvut.cz/uvod/>

^b <http://www.vutbr.cz/portal3/index.php?wapp=library&p4rew=done>

^c <http://www.vkol.cz/cs/>

^d <http://www.zcm.cz/?page=knihovna>

^e <http://knihovna.fpf.slu.cz/>

^f <http://www.bivs.cz/pro-studujici/knihovna-publikace-a-skripta>

^g <http://www.upce.cz/knihovna.html>

3.4 Formáty PSH

Formát zachycující organizační systém vyplývá z účelů jeho využití a podpory softwarových nástrojů. Vzhledem k faktu, že PSH je primárně organizační systém k indexaci dokumentů v katalogu NTK, je jeho hlavním formátem MARC.

3.5 MARC

3.5.1 Historie

MARC (Machine Readable Cataloging, *strojově čitelná katalogizace*) je knihovnický formát určený pro ukládání a výměnu bibliografických záznamů. Byl vytvořen na počátku šedesátých let dvacátého století Henriettou Avram pracující v Kongresové knihovně. MARC se postupem času stal hlavním formátem pro reprezentaci bibliografických záznamů. Knihovny v rámci jednotlivých států jej začaly upravovat pro své lokální potřeby. Jejich nové verze MARCu se tak staly nekompatibilní a kooperace na poli výměny záznamů již nebyla nadále možná. Důvodem byl počet nesourodých formátů a nutnost unikátního transformačního skriptu pro každý z nich.^{14, 15}

Z potřeby jednotného formátu pro sdílení dat vznikl v roce 1977 UNIMARC (druhá verze přišla v roce 1980). UNIMARC byl formát určený pouze pro výměnu dat. Při tomto procesu tedy docházelo ke dvěma transformacím (ze zdrojového formátu do UNIMARCU a z UNIMARCU do formátu výchozího), což velmi ulehčilo život knihovnám, které stavěli své katalogy na záznamech z různých světových zdrojů.

Rozvoj technologií a internetu v 90. letech však změnil způsob distribuce bibliografických záznamů z dávkového na přenos v reálném čase. UNIMARC již tedy nadále nebyl vhodný a knihovníci po celém světě hledali sjednocující řešení. Vzhledem k faktu, že nejrozšířenější modifikací MARCu byl americký USMARC, stal se právě on základem pro nastupující MARC21. Číslovka 21 měla vyjadřovat připravenost formátu na rozvoj techniky jednadvacátého století. MARC21 měl být formátem univerzálnějším a vhodnějším pro mezinárodní kooperaci mezi knihovnami. MARC21 zahrnuje pět hlavních modifikací, každá z nich je určena pro uložení jiného typu dat (*MARC 21 Format for Bibliographic Data, MARC 21 Format for Authority Data, MARC 21 Format for Holdings Data, MARC 21 Format for Classification Data, a MARC 21 Format for Community Information*).¹⁶

V dnešní době je patrně nejrozšířenější formou formátu MARC jeho zápis v XML^a, tzv. MARCXML, který vyvinula a spravuje Kongresová knihovna. Jedná se o klasický XML dokument, který je přepisem MARC polí do XML. Číselné a znakové označení polí a podpolí je představováno atributy u XML elementů.¹⁷

3.5.2 MARC21 pro authority

Základním formátem, ve kterém je PSH uložen, je knihovnický formát MARC21 pro authority.

3.5.3 Struktura formátu

Struktura formátu MARC je implementací mezinárodního standardu ISO 2709^b (*Formát pro výměnu informací*) a ANSI/NISO Z39.2^c (*Bibliografická informační výměna*). Obsah jednotlivých polí je určen číselným označením a je definován samostatně pro každou verzi MARCu.¹⁸

Záznam tvoří tři základní části: leader, kontrolní pole (*controlfield*) a datová pole (*datafield*).

Leader je pole s pevnou délkou 24 znaků, které primárně kóduje některé z informací pro zpracování záznamu. Jedná se o první pole MARC záznamu.

Kontrolní pole jsou pole s informacemi a samotným záznamu. Uchovávají se zde informace o datu vytvoření, osobě, která záznam vytvořila nebo datum poslední modifikace.

^a XML (eXtensible Markup Language, *rozšiřitelný značkovací jazyk*) je obecným značkovacím jazykem, který byl vyvinut konsorciem W3C. Umožňuje vytváření konkrétních značkovacích jazyků pro různé účely. Jeho primárním účelem je ukládání dat, jejich výměna mezi aplikacemi a publikace dokumentů. Pomocí jednotlivých značek nám umožňuje popsat věcně obsah jednotlivých částí dokumentu, jejich vzhled však nijak neovlivňuje. Ten je možné nadefinovat pomocí kaskádových stylů. Dalšími možnostmi jsou transformace do jiného typu dokumentu či jiné XML struktury.¹⁹

^b http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=41319

^c <http://webstore.ansi.org/RecordDetail.aspx?sku=ANSI%2FNISO+Z39.2-1994>

Datová pole obsahují informace o katalogizovaném dokumentu. Skládá se ze základního trojmístného čísla, dvou indikátorů a podpolí specifikovaných jediným znakem.

```
<record xmlns="http://www.loc.gov/MARC21/slim"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.loc.gov/MARC21/slim
http://www.loc.gov/standards/marcxml/schema/MARC21slim.xsd">
  <leader>      nz  a22      n  4500</leader>
  <controlfield tag="001">PSH5706</controlfield>
  <controlfield tag="003">CZ PrSTK</controlfield>
  <controlfield tag="005">20070126161437.0</controlfield>
  <controlfield tag="008">070126na|ann||bab|          |
a|a</controlfield>
  <datafield tag="040" ind1=" " ind2=" ">
    <subfield code="a">ABA013</subfield>
    <subfield code="b">cze</subfield>
  </datafield>
  <datafield tag="150" ind1=" " ind2=" ">
    <subfield code="a">rentgenová difrakční analýza</subfield>
    <subfield code="x">ch</subfield>
  </datafield>
  <datafield tag="550" ind1=" " ind2=" ">
    <subfield code="a">rentgenová analýza materiálu</subfield>
    <subfield code="x">sr</subfield>
  </datafield>
  <datafield tag="550" ind1="9" ind2=" ">
    <subfield code="w">g</subfield>
    <subfield code="a">analytická chemie</subfield>
    <subfield code="x">ch</subfield>
  </datafield>
  <datafield tag="750" ind1="0" ind2="7">
    <subfield code="a">x ray diffraction analysis</subfield>
    <subfield code="2">epsh</subfield>
  </datafield>
</record>
```

Zápis hesla „rentgenová difrakční analýza“ ve formátu MARCXML

3.5.4 MARC v současnosti

Rodina formátů MARC z dnešního pohledu již bohužel neodpovídá požadavkům na jednoduchost manipulace/zpracování počítačem, ani na transparentnost při čtení formátu člověkem. Tyto nedostatky jsou dány stářím MARCu a technologickými možnostmi v době vzniku. Jeho životnost je důsledkem podpory automatizovaných knihovnických systémů, které jej stále využívají jako primární model pro manipulaci se záznamy.²⁰

Základním problémem uložení struktury ve formátu MARC je způsob provázání jednotlivých hesel. Hesla na sebe totiž navzájem odkazují pomocí své textové reprezentace, místo toho, aby pro propojení sloužil jejich unikátní identifikátor (v případě PSH číslo hesla). Z této skutečnosti vyvstává základní problém nutnosti zamezení homonymie hesel v různých kontextech. To znamená, že se v hesláři nesmí vyskytovat dva stejné termíny, přestože svým zařazením patří do odlišných částí struktury a tedy rozdílných vědních oborů.

Některá z polí se mohou vzájemně ovlivňovat a měnit aktuální význam jiného pole v záznamu. Příklad: Význam podpole \$b pole 245 je určen na základě podpole \$a stejného pole. Pokud hodnota podpole \$a končí na znak „rovná se“ (=), obsahuje podpole \$b název dokumentu v cizím jazyce. V případě, že poslední znak podpole \$a je závorka, obsahuje podpole \$b podnázev dokumentu.

Velmi častým jevem je také přidávání dodatečných znaků k aktuálním hodnotám v poli, které přestože jsou nutné k přesné interpretaci záznamu, musí být odstraněny při jeho dalším zpracování.

Z důvodu zastaralosti, složitosti a přetíženosti knihovnická společnost volá po přechodu na nový formát, který by odpovídal dnešním požadavkům na jednodušší počítačové zpracování, lidskou čitelnost, podporu nových standardů i nových metod distribuce dat.^a

^a Velmi známá v knihovnickém světě je iniciativa „*MARC must die!*“ (*MARC musí zemřít!*) založená na stejnojmenném článku z roku 2002 od Roye Tennanta. (http://marc-must-die.info/index.php/Main_Page, <http://www.libraryjournal.com/article/CA250046.html>)

3.6 SKOS

Druhým formátem, který slouží k reprezentaci PSH, je formát SKOS (Simple knowledge organization systém, *jednoduchý systém organizace znalostí*). Do něho byl heslář převeden z autoritních záznamů před necelými dvěma roky za účelem jednodušší manipulace a poskytnutí hesláře veřejnosti. Jednou z inspirací pro jeho implementaci bylo dřívější zveřejnění souborného katalogu Libris Švédské královské knihovny^a. Později se přidala i Kongresová knihovna se svými předmětovými hesly (LCSH).²¹

Simple knowledge organization system je datový model určený pro reprezentaci systémů organizace znalostí jako jsou tezaury a řízené slovníky. SKOS je implementací modelu RDF (*Resource description framework*), který byl vytvořen za účelem zachycení vztahu mezi jednotlivými entitami.²² To znamená, že umí zachytit strukturní vztahy mezi hesly. Hesla na sebe navíc odkazují za použití unikátních identifikátorů (*URI, Unique Resource Identifier*), čímž řeší problém homonymie hesel v různých kontextech. Jak zkratka URI napovídá, těmito identifikátory jsou nejčastěji URL adresy konceptů, globálně dostupné skrze internet. SKOS daným způsobem umožňuje sdílení a linkování hesel na shodné nebo podobné koncepty, nacházející se v dalších organizačních systémech. SKOS se stal 18. srpna 2009 doporučeným standardem W3C^{b c 23}.

^a <http://libris.kb.se/>

^b W3C (World Wide Web Consortium) je mezinárodní komunita, ve které členské organizace a její zaměstnanci společně s veřejností vyvíjejí a ustanovují nové webové standardy. V čele W3C stojí Tim-Berners Lee, vynálezce webu.²⁴

^c <http://www.w3.org/TR/skos-reference/>

3.6.1 Struktura formátu

Pozn. Struktura a možnosti modelu SKOS jsou mnohem rozsáhlejší. Tento stručný popis slouží k demonstraci klíčových funkcí využitých k implementaci PSH v tomto modelu.

SKOS můžeme zapisovat ve formátech, které podporují RDF (nejčastěji XML, Notation3 – N3^a, Turtle^b). PSH je uložený ve formátu SKOS/XML a můžeme s ním nakládat úplně stejně jako s jakýmkoli jiným XML souborem. Vzhledem k tomu, že je implementací RDF modelu, je základní jednotkou jeho struktury trojice - subjekt, predikát a objekt (*RDF triple* - *subject, predicate, object*), která zachycuje vztahy mezi entitami. Subjekt je popisovaný zdroj, objekt je cíl, ke kterému se váže a predikát určuje vztah mezi nimi.

SKOS model je vyjádřen pomocí konceptů strukturovaných vztahy, které vytváří vyšší konceptuální schémata. Jednotlivý koncept je označený unikátním identifikátorem - URI. To umožňuje přiřazování násobných názvů v jakémkoli světovém jazyce, avšak jen jeden termín v každém jazyce může být preferovaným (hlavním) názvem. Ostatním říkáme alternativní nebo nepreferovaná znění, která by měla být s preferovanými synonymní.

Koncepty mohou být spojovány dvěma základními druhy vztahů – vztahy hierarchickými a vztahy asociativními. Vztahy hierarchické vystihují generalitu konceptů, tedy to, že jeden koncept je obecnější než druhý. Vztahy asociativní poukazují na příbuznost prvků, ale nepostulují nadřazenost či vyšší specifitu jednoho z nich.

Vztahy asociativní, na rozdíl od hierarchických neimplikují přenosnost mezi dalšími navázanými koncepty. Ve vztazích hierarchických platí, že když A je

^a Notation3, N3 je zkrácená (ne XML) serializace RDF modelu. Je mnohem kompaktnější a přehlednější než RDF/XML. Byla navržena Tim-Bernersem Lee společně se zástupci komunity sémantického webu.

^b Turtle (Terse RDF Triple Language, *Stručný RDF triple jazyk*) je podmnožinou Notation3. Nemá oficiální status mezi webovými standardy, ale je velmi oblíbený komunitou sémantického webu.

obecnější než B a B je obecnější než C, je A zároveň také obecnější než C. U asociativních tento druh odvozování nelze použít, takže je-li A příbuzné s B a B je příbuzné s C, nemůžeme odvozovat, že by i A bylo příbuzné s C nebo si byly nějakým rysem podobné.^{25 - 27}

```
<skos:Concept rdf:about="http://psh.ntkcz.cz/skos/PSH5706">
  <skos:prefLabel xml:lang="cs">rentgenová difrakční
analýza</skos:prefLabel>
  <skos:prefLabel xml:lang="en">x ray diffraction
analysis</skos:prefLabel>
  <skos:related rdf:resource="http://psh.ntkcz.cz/skos/PSH10660" />
  <skos:broader rdf:resource="http://psh.ntkcz.cz/skos/PSH5653" />
</skos:Concept>
```

Zápis hesla „rentgenová difrakční analýza“ ve formátu SKOS

3.6.2 Mapování hesel na vnější systémy

Spojování konceptů se nemusí omezovat pouze na aktuální schéma. Vztahy mohou existovat mezi koncepty náležících do různých schémat a s různým obsahem. SKOS tak nabízí zachycení propojení do dalších systémů, které mohou být používány v jiných institucích a za jiným účelem. Druh vztahu může být stejně jako uvnitř samostatného schématu hierarchický nebo asociativní. K tomu můžeme určovat konfidenci propojení konceptů na základě míry jejich shody. Můžeme tedy říct, zda jsou dvě hesla stejná (*exact match*, *přesná shoda*) nebo zda jsou si blízká, avšak ne stejná (*close match*, *přibližná shoda*).

Na rozdíl od klasického internetového odkazu, který neposkytuje sémantický popis, předem můžeme tedy s určitou jistotou říci, v jakém jsou propojené strany vztahu. Propojování konceptů linky se sémantickým popisem je základem sémantického webu^a. SKOS byl vytvářen v souladu s touto vizí a v plné míře naplňuje principy *linked data*^{b 28 - 31}.

3.6.3 Výhody SKOS

Implementace PSH/SKOS sebou přináší nesporné výhody. Zápis struktury PSH/SKOS je transparentní, dobře čitelný, neobsahuje duplicitní, ani zbytečné údaje, které nejsou nutné pro jeho využití.

Jak již bylo zmíněno, SKOS pro odkazování mezi hesly využívá unikátních identifikátorů. Odpadá tedy potřeba zamezování homonymních výrazů v různých kontextech. Hesla zůstávají ve svém nejpřirozenějším tvaru bez nutnosti jejich

^a Sémantický web je vize budoucího (částečně naplněná již dnes) webu, kdy popis vztahů mezi jednotlivými entitami umožní počítačům rozumět významu informací a na jejich základě samostatně informace vyvozovat.

^b Linked data (linkovaná data) jsou metodou publikace dat, splňující následující podmínky:

1. Věci identifikujte pomocí URI.
2. Používejte HTTP URI, aby věci byly dohádatelné (člověkem i počítačem).
3. Při odkazování pomocí URI poskytněte popis dané věci např. s využitím standardu RDF/XML.
4. Umožněte objevování nových věcí uváděním linků na příbuzné věci.

úprav pro potřeby formátu. Formát by měl odpovídat potřebám obsahu, ne obsah potřebám formátu.

3.6.4 Aplikace SKOS

Vzhledem k tomu, že se jedná o formát standardní, nebo lépe řečeno standardizovaný, můžeme pro něj najít řadu již vytvořených softwarových aplikací.

Základní aplikací pro práci s jakýmkoli formátem je jeho validátor. Validátor je program, který kontroluje, zda je zápis konkrétního formátu syntakticky v pořádku (*parser*). V případě SKOS následně prověří přítomné vztahy mezi koncepty (jestli mezi dvěma koncepty neexistuje více než jeden vztah, zda samotné odkazované koncepty existují atd.) a zápis podle zvoleného formátu.

V našem případě má však validace ještě jednu nespornou výhodu. Kontrolou souboru s heslářem totiž zjistíme nejen nedostatky v daném formátu, ale i nedostatky zachycené v primárním úložišti, tedy na serveru ve formátu MARC (MARC validátor také existuje, ale kontroluje pouze syntaktickou část, chyby ve vztazích však zůstanou neodhaleny). Validátor tak přeneseně pomáhá odstraňovat chyby, které jsou běžným způsobem dohledatelné jen velmi složitě.

Dalšími programy, které mohou brát slovník jako zdrojový soubor dat, jsou programy vizualizační. Struktura zachycená v datovém modelu může být velmi snadno převedena do své grafické a interaktivní reprezentace. Procházení schématem - katalogem - je v této formě efektnější, „zábavnější“ a také mnohem transparentnější (transparentnost samozřejmě závisí na formě zvolené vizualizace).

Standardní jsou také aplikace, spravující systémy uložené ve formátu SKOS. Po načtení systému je možné modifikovat současné koncepty nebo je přidávat/odstraňovat, stejně jako spravovat veškeré vztahy mezi nimi. Tyto programy nabízí velmi dobré uživatelské rozhraní pro jejich vyhledávání a procházení.

SKOS model se v dnešní době obecně stává velmi frekventovaným při použití pro manipulaci se znalostními systémy a podporuje jej většina aplikací určených pro

práci s nimi. Často bývá označován za nejlepší alternativu při převodu znalostního systému do sémantického formátu.^{32, 33}

3.7 Prohlížení hesláře

Na stránkách NTK se v sekci PSH pod odkazem „*Prohlížení hesláře*“^a nachází grafická reprezentace jeho struktury. Představuje alternativní přístup ke knižnímu fondu NTK. Nabízí logický způsob procházení sbírky na základě tematického zaměření po hierarchické a asociativní ose.

Na hlavní stránce se nachází seznam 44 hlavních tematických větví. Po přechodu na konkrétní heslo se zobrazí jeho grafická reprezentace (obr. 3.2). Je tvořena středovým kruhem s aktuálním heslem a okolními vazbami na heslo nadřazené, podřazená a příbuzná (obr. 3.3). Jejich vztah je indikován malou ikonou u každého z nich. V levém sloupci se zobrazují nepreferovaná znění v češtině i angličtině. Do anglického jazyka je možné přepnout i celé prohlížení hesláře (obr. 3.4).

Pod seznamem podřazených hesel na levé straně se nacházejí odkazy do příbuzných datových množin. V případě hesláře se jedná propojení na hesla DBPedia^b a Předmětová hesla Kongresové knihovny^c. K jejich namapování dochází podle anglického preferovaného i nepreferovaného znění. Ze dvou formátů, ve kterých je heslář uložen, jsou tyto vazby zaznamenány pouze ve formátu SKOS^{d,e}. V případě preferovaných znění je vazba popsána jako „přesná shoda“^f, v případě

^a <http://www.techlib.cz/cs/katalogy-a-database/psh/prohlizeni-psh/>

^b DBPedia je databáze strukturovaných dat extrahovaných z Wikipedie zpřístupněná pro web. Strukturované informace umožňují uživateli procházet datové závislosti a mapovat hesla na další datové množiny.

^c Předmětová hesla Kongresové knihovny (Library of Congress Subject Headings, LCSH) je tezaurus předmětových hesel spravovaný Kongresovou knihovnou. Zde hraje stejnou roli jako Polytematický strukturovaný heslář v Národní technické knihovně. Jedná se o jeden z nejznámějších knihovnických systému organizace znalostí.

^d Formát MARC nenabízí pole pro jejich uložení.

^e Reprezentaci hesla v tomto formátu je možné zobrazit po kliknutí na odkaz „Zobrazení hesla ve formátu SKOS“. Reprezentaci hesla ve formátu MARC se nachází u autoritního záznamu hesla v katalogu NTK.

^f Element *skos:exactMatch*.

nepreferovaných jako „blízká shoda“^a. Pro odkazování na PSH slouží URI hesel ve tvaru http://psh.ntkcz.cz/skos/id_hesla (např. <http://psh.ntkcz.cz/skos/PSH67>).

Pod grafickou reprezentací se simultánně zobrazují dokumenty z katalogu Národní technické knihovny a Národní knihovny ČR (NK ČR) označené aktuálním heslem. Vyhledání dokumentů z katalogu NK ČR bylo docíleno automatickým namapováním hesel PSH na Věcné autority NK ČR^b. Výsledek se tedy zobrazí jen v případě shodných hesel v obou znalostních systémech. Z nalezených dokumentů vede link přímo do katalogu, kde je možná jeho výpůjčka nebo rezervace.

Česky English
 Hledat heslo



anorganická chemie

URI: <http://psh.ntkcz.cz/skos/PSH5740>^U

Anglický ekvivalent: inorganic chemistry

Příbuzná hesla:

- chemické sloučeniny
- farmaceutická chemie
- struktura pevných látek

Nadřazené heslo:

- chemie

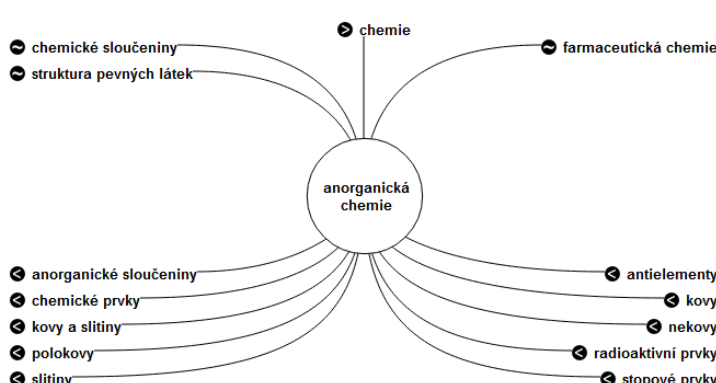
Podřazená hesla:

- anorganické sloučeniny
- antielementy
- chemické prvky
- kovy
- kovy a slitiny
- nekovy
- polokovy
- radioaktivní prvky
- slitiny
- stopové prvky

[Zobrazení hesla ve formátu SKOS](#)^U

[Chybí mi tu — návrh nového hesla](#)^U

Zobrazit tag z řízeného hesla

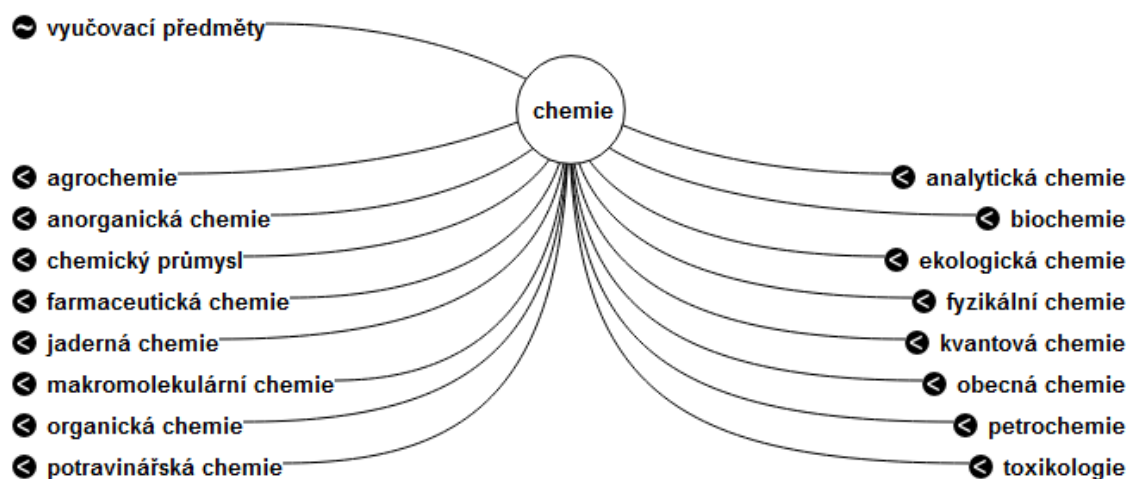


Katalog Národní technické knihovny

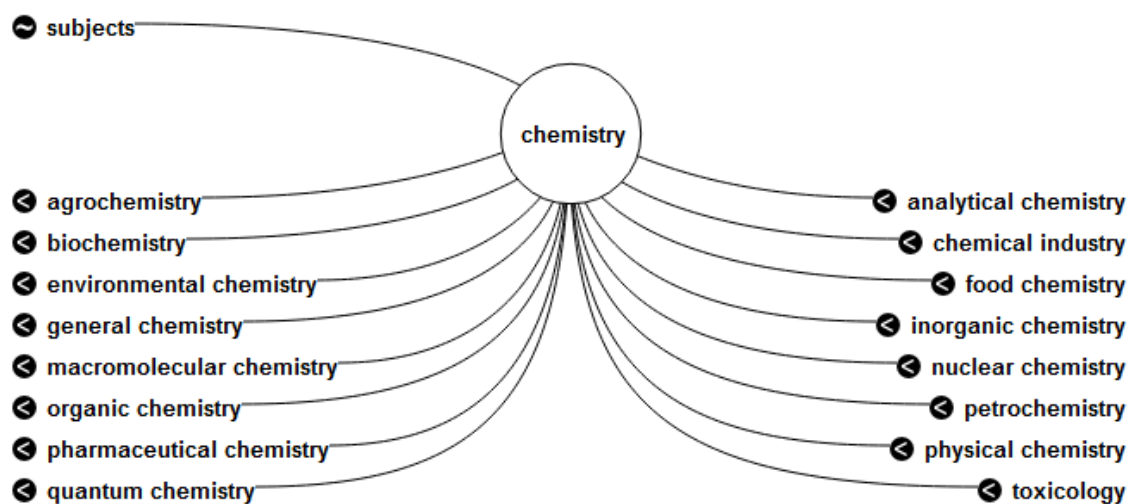
Obr. 3.2 Rozhraní pro prohlížení PSH

^a Element *skos:closeMatch*.

^b Věcné autority Národní knihovny České republiky patří do souboru Národních autorit ČR spravovaných Národní knihovnou. Jedná se o největší systém věcných autorit v ČR, který v současné době zahrnuje přes 200 000 hesel. Na jeho tvorbě se podílí většina paměťových institucí, v důsledku sdílení některých záznamů z NK ČR.³⁴



Obr. 3.3 Grafická reprezentace hesla chemie v českém jazyce



Obr. 3.4 Grafická reprezentace hesla chemie (chemistry) v anglickém jazyce

3.7.1 Vytváření úryvků metadat s hesly PSH

Další funkcí zpřístupněnou na stránkách prohlížení hesláře je vytváření úryvků metadat^a s hesly PSH, které slouží k předmětovému popisu HTML stránek (obr. 3.5). Metadata je možné generovat ve dvou nabízených formátech - *Dublin Core* a *Common Tag*. Jejich vložení do HTML dokumentu dosáhneme jeho sémantického popisu.³⁵

„**Dublin Core** je soubor metadatových prvků, jehož záměrem je usnadnit vyhledávání elektronických zdrojů. Původně byl vytvořen jako popis zdrojů na WWW sestavený přímo autorem, postupně ale zaujal instituce zabývající se formálním zpracováním zdrojů, jako jsou muzea, knihovny, vládní agentury a komerční organizace.“ [DUBLIN CORE, 2011] V současné době se jedná patrně o nejrozšířenější způsob zápisu předmětových metadat na webu.³⁶

„**Common Tag** je poměrně mladý sémantický formát pro tagování webového obsahu, přispívající k jeho vyšší provázanosti, vyhledatelnosti a uspořádanosti.“ [COMMON TAG, 2009] Byl představen v červnu roku 2009 pod záštitou společnosti Yahoo!, velmi brzy jej však začali podporovat i další společnosti nabízející internetové služby (*Freebase^b*, *Zemanta^c*, *Zigtag^d* aj.).³⁷

^a Metadata jsou strukturovaná data o datech. Bibliografický záznam je jejich dokonalým příkladem.

^b <http://www.freebase.com/>

^c <http://www.zemanta.com/>

^d <http://www.zigtag.com/>

Níže uvedené úryvky metadat lze vložit do HTML dokumentu pro dosažení jeho sémantického popisu:

Dublin Core:

```
<p about="" xmlns:dc="http://purl.org/dc/elements/1.1/">  
  <a href="http://psh.ntkcz.cz/skos/PSH5579" rel="dc:subject">iontová vazba</a>  
</p>
```

CommonTag:

```
<body xmlns:ctag="http://commontag.org/ns#" rel="ctag:tagged">  
  <span typeof="ctag:Tag" rel="ctag:means" resource="http://psh.ntkcz.cz/skos/PSH5579"  
    property="ctag:label" content="iontová vazba" />  
</body>
```

Obr. 3.5 Vytvořené úryvky metadat pro heslo „iontová vazba“

4 Cíl práce

PSH je v České republice ojedinělým produktem, který poskytuje rozsáhlé možnosti na poli kategorizace, strukturního propojení příbuzných datových množin, vizualizace katalogů a mnoho dalších. Za tímto účelem však musí být aktuální, sledovat současné trendy, své uživatele, a vyhledávat projekty, které by mohli z jeho vlastností těžit.

PSH v současné době nemá standardizované procesy správy, především vlastní aktualizace, ani zdroje, ze kterých by periodicky čerpal. Jediným aktuálním zdrojem je formulář pro návrh nového hesla, který ovšem není moc využíván. Běžní uživatelé ve většině případů neví, co to Polytematický strukturovaný heslář je, natož aby navrhovali hesla, která by se do něj měla zařadit.

Je nutné implementovat funkce, které budou zpracovávat statistiky z nejběžnějších zdrojů, které jsou využívány po celém světě. Těmi jsou především klíčová slova bibliografických záznamů a dotazy uživatelů při vyhledávání v katalogu. K aktualizacím musí docházet pravidelně/průběžně a pokud možno, co nejvíce automatizovaně, tak aby uživatel (především aktivní - katalogizátor), používal vždy všechna dostupná hesla.

Využití PSH velmi úzce souvisí s nástroji pro jeho prezentaci. Různí uživatelé však mají různé požadavky na rozhraní. Uživatelské rozhraní musí být těmto rozdílným skupinám (aktivní a pasivní uživatelé) přizpůsobeno.

Aktivní uživatelé vyžadují primárně přehledný pohled na strukturu hesel i na popis jednotlivých konceptů. Současný nástroj pro prohlížení PSH – *PSH Manager* - je zastaralý a nekompatibilní s aktuálním heslářem. Dostupná webová služba je aktuální, ale nesplňuje požadavky na přehlednost a rychlost práce. Výsledný software by měl poskytovat požadované funkce tak, aby práce byla intuitivní, komfortní, a tedy i rychlá a efektivní.

Běžní uživatelé naopak očekávají samopopisnou prezentaci, která graficky zaujme (je efektní), obsahující přidanou hodnotu v podobě dodatečných funkcí. Ani zde není online prohlížení PSH plně dostačující z pohledu nedořešeného vyhledávání a

rozvržení stránky, skrývající právě dodatečné funkce, které nejsou na první pohled viditelné a zřejmé.

Primárním účelem hesláře je předmětová kategorizace dokumentů při jejich zařazování do katalogu. Možností využití je však více a dalo by se říci, že závisí především na uživateli. Je velmi důležité, z jakého úhlu na celou věc pohlížíme. Základní otázkou v tomto směru tedy není jen „co“, ale také „jak“ a co nám konkrétní implementace může přinést.³⁸

5 Užité programovací technologie

Následující část obsahuje popis programovacích technologií a nástrojů, které byly využity při realizaci projektů spojených se správou PSH.

5.1 Python

Python je interpretovaný objektově orientovaný jazyk. Byl navržen v roce 1990 nizozemským programátorem Guido van Rossumem. Jeho distribuce je nabízena zdarma pro většinu běžných platforem (Windows, Linux, Mac OS), a to jak v podobě zdrojových kódů, tak instalačního balíčku. Python je někdy označován jako jazyk hybridní (paradigmatický), při psaní programů totiž umožňuje využívat nejen objektově orientovaný přístup, ale také procedurální a v jisté míře i funkcionální. Je velmi jednoduchý z hlediska učení. V dnešní době bývá považován za jeden z nejlepších programovacích jazyků pro začátečníky[18]. Důležitým aspektem je vysoká produktivita (rychlost) při psaní kódu, umožněná stručností jeho zápisu a také velkým množstvím již vytvořených programových modulů.

V současné době existují dvě hlavní stabilní verze Pythonu – 2.7 a 3.2. Rozdílnosti mezi 2.x a 3.x větvemi zapříčiňují vzájemnou nekompatibilitu programů. Většina programových modulů však dosud nebyla aktualizována pro funkcionalitu s 3.x verzí. Všechny programy popsané v této práci využívají Python ver. 2.5 a vyšší.^{39, 40}

5.2 Django Framework

Django je open source webový aplikační framework^a napsaný v Pythonu, který se volně drží architektury model-view-controller^b (MVC). Původně bylo navrženo pro správu zpravodajsky orientovaných stránek společnosti The World Company v Lawrence, Kansas, později, v červnu 2005, bylo vydáno veřejně pod open-sourceovou licencí. Framework byl pojmenován po jazzovém kytaristovi Django Reinhardtovi.^{41, 42}

^a Programovací framework představuje soubor funkcí ulehčujících práci v daném programovacím jazyce. Webový aplikační framework umožňuje jednodušší tvorbu dynamických webových stránek, aplikací a služeb.

^b MVC (model-view-controller, *model-pohled-kontroler*) je softwarová architektura, která rozděluje datový model aplikace, uživatelské rozhraní a řídicí logiku do tří nezávislých komponent tak, že modifikace některé z nich má minimální vliv na ostatní.⁴³

5.3 JavaScript

JavaScript je multiplatformní, objektově orientovaný skriptovací jazyk. Primárně se používá jako interpretovaný programovací jazyk pro tvorbu WWW stránek, konkrétně k implementaci interaktivních prvků stránky a pokročilých grafických funkcí. Na rozdíl od většiny interpretovaných jazyků se jeho kód spouští až na straně klienta (client side scripting) ve webovém prohlížeči, z čehož plynou jistá bezpečnostní omezení, například pro práci s lokálními soubory.⁴⁴

5.4 jQuery

jQuery je volně stažitelná, open source knihovna programovacího jazyka JavaScript, která zjednodušuje úkony při programování webových stránek. Mezi hlavní vlastnosti patří zjednodušení manipulace s DOM^a objekty, řízení funkcí (event handling), využití AJAXu^b a tvorba animací v rámci webových stránek. jQuery celkově zjednodušuje syntaxi JavaScriptu a stará se o většinu problémů způsobených různými implementacemi chování HTML + CSS + JavaScript dokumentů ve všech hlavních internetových prohlížečích (Internet Explorer, Mozilla Firefox, Opera, Google Chrome a Safari). jQuery bylo představeno v roce 2006 a od té doby se stalo nejrozšířenější používanou JavaScriptovou knihovnou.⁴⁵

5.5 MySQL

MySQL je multiplatformní databáze. Komunikace s ní probíhá pomocí jazyka SQL^c. Podobně jako u ostatních SQL databází se jedná o dialekt tohoto jazyka s některými rozšířeními. Je volně stažitelná a velmi rozšířená, kvůli své snadné implementovatelnosti i výkonu.⁴⁶

^a DOM (Document Object Model, *objektový model dokumentu*) je objektově orientovaná reprezentace XML nebo HTML dokumentu. DOM je API umožňující přístup či modifikaci obsahu, struktury, nebo stylu dokumentu, či jeho částí.

^b AJAX (Asynchronous Javascript and XML, *asynchronní JavaScript a XML*) je obecné označení pro technologie vývoje interaktivních webových aplikací, které mění obsah svých stránek bez nutnosti jejich znovunačtení.

^c SQL (Structured Query Language, *strukturovaný dotazovací jazyk*) je standardizovaný dotazovací jazyk používaný pro práci s daty v relačních databázích.

5.6 HTML

HTML (HyperText Markup Language, *hypertextový značkový jazyk*) je dominantním značkovacím jazykem pro publikaci stránek na internetu. Je aplikací dříve vyvinutého rozsáhlého univerzálního značkovacího jazyka SGML^a. Je charakterizován množstvím předdefinovaných značek (elementů) a atributů. S jejich pomocí formátuje rozložení stránky, textu i všech ostatních objektů.⁴⁷

5.7 CSS

CSS (*Cascading Style Sheets, tabulky kaskádových stylů*) je jazyk pro popis způsobu zobrazení stránek napsaných v jazycích HTML, XHTML nebo XML. Jeho hlavním smyslem je oddělení vzhledu dokumentu od obsahu. Jeho možnosti formátování webových stránek jsou mnohem vyšší než u samotného jazyka HTML. Jeho poslední verze, CSS3, implementuje řadu funkcí - efektů, které dříve byly možné pouze s využitím jiných programovacích jazyků, nejčastěji JavaScriptu nebo PHP.⁴⁸

^a SGML (Standard Generalized Markup Language, *standardní zobecněný značkový jazyk*) je univerzální značkový metajazyk, který umožňuje definovat značkovací jazyky jako své vlastní podmnožiny.

6 Procesy správy hesláře

Jak již bylo řečeno dříve, správa znalostních systémů by se dala rozdělit na dva hlavní body – aktualizaci systému a jeho opravy. U PSH tomu není jinak. Revize hesláře plně vychází z procesů jeho aktualizace.

6.1 Aktualizace PSH

Do současnosti docházelo k zařazování nových hesel spíše nárazově. Často byli oslovováni externí odborníci v konkrétním oboru k provedení revize jen určité části hesláře. S tím samozřejmě přicházel problém nekonzistence PSH v různých oborech podle subjektivního a komplexního pohledu každého jedince. Zejména z tohoto důvodu nejsou všechny tematické větve stejně strukturované a detailní. Tato nesourodost vychází především malé znalosti celého systému a vytržení revidovaného tématu z kontextu hesláře. V současnosti již nedochází k oslovování s externích pracovníků. Referát PSH ale v záležitostech aktualizace velmi intenzivně spolupracuje s Oddělením správy autoritních souborů NTK.

Přestože již v současné době nezařazují hesla experti v daných oborech, ale pracovníci, kteří mají svým vzděláním velmi často daleko ke konkrétnímu tématu, je dle mého názoru, využitím informačních kanálů, spoluprací a výslednou koncesí, tento přístup více než dostatečnou náhradou. O nových heslech či změnách struktury nikdy nerozhoduje pouze jeden člověk.

6.2 Zdroje aktualizací

Zdroje aktualizací PSH bychom mohli rozdělit na interní a externí. Interní zdroje vycházejí z dat vytvořených v rámci knihovny (nejčastěji katalogizátory). Externí představují data vytvořená uživateli.

6.3 Formulář pro návrh nového hesla

Nejpřímější cestou při návrhu nových hesel je formulář pro návrh nového hesla^a zpřístupněných na stránkách PSH (obr. 6.1). Je určený pro všechny uživatele hesláře. Jediný povinný údaj je samotné navrhované heslo. Ostatní jako ekvivalentní hesla, kontakt na uživatele a poznámka jsou nepovinné.

^a <http://www.techlib.cz/cs/614-navrh-noveho-hesla>

Jak by se dalo očekávat, tento formulář není běžnými uživateli moc často využíván. To však neplatí ani pro katalogizátory, kteří by jej vzhledem k svému postavení a vlastnímu zájmu používat měli. Tato situace nás nutí spoléhat více na možnosti analýzy nepřímých zdrojů, kterými jsou v našem případě klíčová slova a logy vyhledávání.

Návrh nového hesla PSH

Nové heslo PSH	
Synonymum nového hesla (česky)	
Nové anglické heslo PSH	
Synonymum nového hesla (anglicky)	
E-mail	
ISBN nebo ISSN	
Poznámka	

Obr. 6.1 Formulář pro návrh nového hesla

6.4 Analýza klíčových slov

Klíčová slova jsou v praxi uživatelsky vytvářená předmětová hesla. Vytváří se nejen v knihovnách při popisu dokumentů, ale vytváříme je my všichni např. při třídění elektronické pošty, při tagování^a internetového obsahu, v RSS^b čtečce^c ⁴⁹,

^a Tagování (od anglického slova tag - značka) je název pro přiřazování štítků (klíčových slov) internetovému obsahu.

^b RSS (Really Simple Syndication, *opravdu jednoduchá syndikace*) představuje rodinu XML formátů určených pro čtení novinek na webových stránkách. RSS umožňuje uživateli přihlášení k odběru nových článků na stránkách, které tuto službu poskytují (RSS zdroj, RSS kanál).

^c RSS čtečka je aplikace pro spravování přihlášených RSS zdrojů.

v záložkách, při používání speciálních služeb pro vytváření záložek (delicio.us^a, digg.com^b aj.), při přiřazování hashtagů^c na Twitteru^d a dalších organizačně-třídících procesech.

Analýza klíčových slov představuje jednu z hlavních metod aktualizace hesláře (systémů organizace znalostí obecně⁵⁰) z interních knihovních zdrojů. Jedná se o zdroj nepřímý, protože indexátoři by měli být těmi, kdo upozorňuje na hesla, která mezi předmětovými chybí. Právě tento faktor klíčová slova odráží.

Klíčová slova se u bibliografických záznamů ve formátu MARC nacházejí v poli 653 (předmětová hesla v poli 650). Je zapotřebí extrahovat obsah tohoto pole a vytvořit jednoduchou statistiku, která zachycuje počet výskytů jednotlivých termínů.

Hlavní záhlaví	Truesdell, Cliff, 1967-
Název	Introducing Reason 4 / Cliff Truesdell
Naklad. údaje	Indianapolis : Wiley, c2008
Popis (rozsah)	xx, 435 s. : il. + 1 CD-ROM
Poznámka	Obsahuje rejstřík
Produkt	Reason 4
Klíčové slovo	hudební software
	počítačová kompozice hudby
Heslo PSH	aplikační programy -- vt
	elektronická hudba -- um
	digitální záznamy -- el

kandidáti

stávající
hesla PSH

Obr. 6.2 Ukázka bibliografického záznamu

^a <http://www.delicious.com/>

^b <http://digg.com/>

^c Hashtag (znak hash - #, tag - značka) je způsob štítkování zpráv na mikroblogovací službě Twitter. Uživatelé označují své zprávy pomocí tagů ve formě „#tag“. Tyto štítky umožňují následné filtrování zpráv podle obsahu.

^d <http://twitter.com/>

Jako vstupní formát byl administrátorem Alephu vygenerován soubor obsahující jedno klíčové slovo na řádku pro všechny záznamy, které jej obsahují. Po jeho zpracování (normalizace termínů, sečtení stejných termínů) vznikl nový soubor s dvojicemi hodnot – „klíčové slovo“ : „počet výskytů“, s prahovou hodnotou multiplicity čtyři. Výsledkem bylo 810 posuzovaných kandidátů na nová předmětová hesla. Vyhodnocování vhodnosti hesel probíhalo manuálně v úzké spolupráci s Oddělením správy autoritních souborů.

	A	B	C	D	E	F
1	Počet		Termín			
2	180	ev	podnikové informační systémy	D	13633	management information systems
3	180		webové stránky	N		
4	167	vt	3d	D	13622	three dimensional
5	159		síťově orientované programovací jazyky			
6	156	sp	modelářství	D	13624	modeling
7	126		webové aplikace	N		
8	125		informační technologie	N		
9	117	vt	agenti	D	13625	agents
10	115		multiagentní systémy	N		
11	113		komunikační technologie	N		
12	99	if	bioinformatika	D	13632	bioinformatics
13	90	um	digitální fotografie	D	13631	digital photography
14	89	vt	zvukové editory	D	13630	audio editors
15	89	if	webové programování	D	13635	
16	87	if	webové inženýrství	D	13634	web engineering
17	83		grid computing	N		
18	82	vt	sémantický web	D	13629	semantic web
19	79		dynamické webové stránky	N		
20	76		world wide web	N		
21	69		bezdrátové komunikace			
22	68		peer-to-peer computing	N		
23	61	vt	počítačové vidění	D	13621	computer vision
24	60	if	genetické algoritmy	D	13628	genetic algorithms
25	60		dálková ovládání			
26	59		bitmapové editory	N		
27	58	vt	interakce člověk - počítač	D	13636	human computer - interaction
28	58	if	evoluční algoritmy	D	13627	evolutionary algorithms
29	58		nízkoenergetické domy	N		
30	57		mobilní komunikační sítě	N		
31	56		videostřižna	N		
32	54	vt	2d	D	13622	two dimensional

Obr. 6.3 Sdílená tabulka pro vyhodnocování termínů

6.4.1 Výsledky analýzy klíčových slov

Výsledkem analýzy klíčových slov bylo celkem 114 nových deskriptorů se svými ekvivalenty a 119 nedeskriptorů. Většina zamítnutí ostatních kandidátů pramenila z přílišné specifiky nebo naopak neurčitosti hesla. Některé termíny byly nevhodné už ze své podstaty a správně by se neměly vyskytovat ani mezi klíčovými slovy (geografické názvy, názvy produktů).

Klíčová slova s vyšší než prahovou multiplicitou výskytu (4)	810 (100%)
Nové deskriptory	114 (14%)
Nové nedeskriptory	119 (15%)
Počet zamítnutých	577 (71%)

Tabulka 6.1 Výsledky analýzy klíčových slov

6.4.2 Následky analýzy klíčových slov

Analýza pole 653 však sebou nepřinesla jen pozitivní výsledky ve formě rozšíření hesláře o nové termíny, ale také otázku, co s bibliografickými záznamy, které obsahují vybraná klíčová slova. Termíny zařazené do hesláře musí být samozřejmě z pole 653 smazány, kvůli výsledkům vyhledávání podle klíčových slov nebo hesel PSH. Na druhou stranu, po jejich smazání konkrétní záznamy (v řádu tisíců) ztratí jeden ze svých předmětových deskriptorů.

Nejjednodušší možností je tento problém neřešit a se situací se spokojit. To však způsobí, že výsledek dotazu uživatele na konkrétní termín nedosáhne zdaleka takové kvality - relevance, jaké by měl.

V případě, že by knihovnický systém Aleph disponoval funkcemi pro dávkovou změnu pole 653 na pole 650, stačilo by na vstupu vložit seznam termínů v konkrétním formátu. Tuto funkci však (podle administrátora systému) Aleph neobsahuje a teoreticky zbývá poslední možnost, provést všechny změny ručně, neboť není k dispozici ani API, které by ulehčilo funkci implementovat.

Nakonec nás však napadla ještě jedna možnost: „Když nemůžeme automatizovat systém, nemohli bychom automatizovat samotnou lidskou práci?“ Ukázalo se, že to možné je a odpovědí na tuto otázku byl softwarový balík Sikuli.

6.4.3 Sikuli

Sikuli je programovací nástroj sloužící k automatizaci grafického uživatelského rozhraní. Umožňuje vyhledávat vzory na grafickém výstupu a automatizovat uživatelskou interakci. Využívá výřezy plochy (*screenshoty*) jako vizuální reference přímo v programovém kódu.⁵¹ Ten je, dle mého názoru, intuitivní a umožňuje používat příkazy jako: klikni na dané místo (*click*) nebo napiš text (*type*). Jeho syntaxe je jednoduchá, je shodná s jazykem Python. To z důvodu využití jazyka Jython, který je implementací Pythonu v Javě.

6.4.4 Programování v Sikuli

Využití obrázků jako přímých vizuálních referencí znamená skutečné použití příkazu „klikni na tento obrázek, který reprezentuje část mé plochy“ (obr 6.4).

Například:

click()

Obr. 6.4 Ukázka programového kódu v Sikuli IDE

Takto může vypadat klasický programový kód vytvořený v Sikuli IDE^a. Sikuli IDE je textový editor, dodávaný jako součást balíku Sikuli. Poskytuje funkcionalitu k pohodlnému pořizování výřezů plochy, které vkládá přímo na místo textového kurzoru. Na konkrétní místo nemusíme pouze klikat, ale obrázek je možné použít například k upřesnění polohy na ploše nebo jako součást podmínky při běhu programu.

Způsob vyhledání vzoru na ploše záleží na jeho velikosti. Pro menší (ikona, tlačítko) je využita metoda normalizované křížové korelace, která povoluje částečné změny grafických vlastností obrázku (kontrast, světlost).⁵² Pro velké vzory byl adaptován algoritmus založený na metodě SIFT (Scale-invariant feature transformation). Ten se dříve používal k rozeznávání postav v pouličním záběru.

^a IDE (Integrated Development Environment, *integrované vývojové prostředí*)

Obraz je popsán systémem vektorů určených klíčovými body. Tím umožňuje částečné změny úhlu pohledu nebo změny rozlišení.^{53 - 55}

6.4.5 Výhody vizuálního přístupu

Hlavní výhodou Sikuli oproti běžným postupům je právě jeho nekonvenčnost, kvůli které ho můžeme využít tam, kde běžné metody selhávají. Sikuli dokáže automatizovat prakticky jakýkoli software s grafickým uživatelským rozhraním. Automatizace GUI je navíc ve většině případů méně intelektuálně náročná, protože je transparentní - používáme příkazy, které jen zastupují úkony, které bychom provedli při běžné práci.

6.4.6 Nevýhody vizuálního přístupu

Vizuální přístup k automatizaci grafického prostředí s sebou přináší řadu omezení, která vyplývají z jeho samotné podstaty. Ve srovnání s automatizací (programováním) na nižší úrovni je mnohonásobně pomalejší^a a neumožňuje současnou práci uživatele na PC. Jedná pouze o parametrické nedostatky, ale existují i další překážky, které mohou zabránit v jeho využití.

6.4.6.1 Vzhled systému

Jedním z problémů obrazového přístupu je nutnost zachování vzhledu operačního systému nebo programu, který automatizujeme. Ze stejného důvodu je také velmi často nemožné použít program napsaný v Sikuli na jiném než zdrojovém počítači. Jedním z řešení by mohlo být například vynucení původního nastavení vzhledu systému. To však jen v případě, že se jedná o stejné operační systémy.

6.4.6.2 Nestandardní chování hardware a software

Při běhu programu je nutné brát v potaz také omezení vyplývající z hardwarové konfigurace počítače. Pomalý procesor nebo malé množství operační paměti může způsobovat neplynulý běh systému. Sikuli nenalezne vzor, kterému trvá delší dobu než se objeví na obrazovce a program skončí v lepším případě chybou. V horším případě se exekuce kódu nezastaví, ale bude se provádět na úplně jiném místě, než bylo naplánováno. Z tohoto důvodu je nezbytné, aby program byl robustní – to znamená, byl na tyto situace připraven.

^a I přesto je Sikuli stále mnohem rychlejší a přesnější než lidská práce.

K nekonzistentnímu chování může docházet i na straně software. K přerušení programu může dojít například při aktualizaci programů a objevení vyskakovacího okna. To může nejen překrýt potřebnou část plochy, ale především mění aktuální pracovní (aktivní) vrstvu, se kterou Sikuli pracuje.

6.4.7 Automatizace Aleph GUI

Jako vstupní datová struktura byl zvolen seznam tuplů^a (trojic pro preferovaná znění, čtveřic pro nepreferovaná znění). Každý se skládal z hlavního termínu, názvu souboru určeného k uložení editovaných bibliografických záznamů a ze zkratky větve PSH, do které byl termín zařazen. Pro nepreferované znění bylo nutné navíc uvést deskriptor, ke kterému se váže.

Samotný program představoval for-cyklos^b, který postupně iteroval přes všechny termíny a prováděl operace v grafickém rozhraní softwaru Aleph.

Program zahrnoval následující činnosti:

1. vyhledání bibliografických záznamů s aktuálním heslem v poli 653
2. uložení výsledku hledání do souboru
3. vyplnění formuláře *Globální změny* a jeho použití na záznamy uložené v souboru

Výsledný program (délka čtyřicet řádků) se omezuje jen na použití čtyř příkazů a konstrukci for-cyklu. Čas strávený nad jeho psaním byl značně kratší, než jaký by byl při manuálním provádění. Skript je navíc znovupoužitelný a lehce upravitelný pro další potřeby.

6.4.8 Použitá řešení

I přes možnosti klikání na konkrétní místa na ploše je stále nejlepším řešením automatizace vstupu z klávesnice. Tato metoda je rychlejší a hlavně přesnější. Využití myši by mělo přicházet až v případech, že akci není možné klávesnicí nahradit.

^a Tuple je n-tice elementů v určeném pořadí.

^b For-cyklus – jedna ze základních řídicích programovacích struktur, která provede daný sled příkazů nad každým z určené množiny prvků

Při vyplňování polí je nutné brát zřetel na aktuální rozdílnost funkcí „napiš“ (type) a „vlož“ (paste). Funkce vložení nyní funguje jako dočasné řešení při psaní mezinárodních znaků nebo s jiným rozložením klávesnice než US-QWERTY.

6.4.9 Výsledky automatizace

Automatizace systému Aleph se v praxi osvědčila. Posloužila svému hlavnímu účelu - reindexaci bibliografických záznamů, a to v kratším čase a bez zbytečného zapojení lidské práce. Navíc ukázala možnosti tohoto nástroje při automatizaci dalších úkonů, které jsou nejen v knihovnickém prostředí na denním pořádku (kontrola a zadávání dat).

6.5 Analýza LOGu vyhledávání v katalogu NTK

Celý dnešní web se zabývá primárně uživatelem, zkoumáním jeho požadavků, zálib a potřeb. Jako každá jiná firma se i knihovna snaží o to, aby svým uživatelům nabízela informace, které hledají. Z tohoto důvodu je nutné používat jazyk společný oběma „komunikujícím“ stranám, který se samozřejmě v průběhu času mění.⁵⁶ Některé termíny zastarávají a nahrazují je jiné, některé jsou nové a přicházejí se svou dobou.^a V každém případě se vždy musí jednat o firmu – instituci, která se podřídí požadavkům a vyjde svému zákazníkovi vstříc.

6.5.1 Zdroj informací

Zdrojem informací již dnes není komunikace mezi lidmi, ale hlavně komunikace mezi člověkem a vyhledávacím rozhraním. Jak mnohé studie ukázali, drtivá většina uživatelů svůj dotaz dále nespecifikuje, nepoužívá „složitě“ pokročilé vyhledávání a volí nejjednodušší cestu vyhledávání ve všech polích.^{57, 58} Většinou se jedná o zvyklost z běžných internetových vyhledávačů, jindy odradí až příliš mnoho vyhledávacích možností. Každopádně, výsledek je závislý vždy jen na popisu dokumentů, ne na jejich plném textu, který knihovna většinou ani v elektronické podobě nemá. To v překladu znamená, že pokud hledáme určitý termín, výsledkem budou vždy jen publikace, které mají tento termín v názvu nebo v některém z polí předmětového popisu.^b Je tedy více než důležité, aby knihovny věděli, co uživatelé hledají a jaké deskriptory mají standardně přiřazovat.

6.5.2 Analýza logů

Analýza logů je dnes již klasickým způsobem vedoucím k porozumění chování uživatelů při vyhledávání informací.^{59 - 62} Patrně nejlepším příkladem jsou

^a To platí především pro rychle se rozvíjející technické obory.

^b Například při vyhledávání slova „fyzika“ rozhodně nenajdeme všechny články spadající do tohoto oboru, ale jen ty, které mají slovo „fyzika“ v názvu, nebo deskriptor „fyzika“. Bude se jednat jen o zlomek dokumentů relevantních k tomuto dotazu.

statistiky vyhledávání Googlu^a, které slouží od personalizace reklamních nabídek až po predikci nadcházející chřipkové epidemie^b.

Log se říká záznamu nebo souboru záznamů, do kterého program ukládá informace o svém běhu (procesy, informace o uživateli, chybové zprávy). Tyto soubory následně slouží k zpětné analýze programu, optimalizaci, řešení chyb nebo v našem případě porozumění chování uživatelů při vyhledávání. Log nemá standardizovanou podobu, každý programátor jej implementuje podle svého, pokud vůbec. V logu jsou velmi často zaznamenány nejen vyhledávaná slovní spojení a parametry vyhledávání, ale i další informace o uživateli jako IP adresa^c nebo datum a čas.

Některé z informačních systémů v sobě již mají integrované funkce pro analýzu uživatelských dotazů a vytváření statistik. Většinou však instituce musí vytvořit vlastní metody vhodné pro jejich systém a jeho použití.

6.5.3 Log knihovnického informačního systému Aleph

Log systému Aleph se vytváří vždy při startu programu a informace se do něj ukládají až do té doby, než je nucen svou činnost ukončit^d. Vzhledem k množství informací a délce času zaznamenávání jeho velikost může dosáhnout až řádu jednotek gigabytů. Avšak jen velmi malý zlomek těchto dat je pro nás užitečný. Informace pro nás relevantní jsou vyhledávací řetězec s parametry, IP adresa uživatele a čas vyhledávání.

^a <http://www.google.com/trends>

^b Google Flu Trends (*chřipkové trendy*) na základě vyhledávání výrazů úzce spojených s chřipkou, jejími příznaky a léčením dokáže poměrně spolehlivě předpovědět nadcházející chřipkovou epidemii. (<http://www.google.org/flutrends/>)

^c IP adresa (Internet Protocol, *internetový protokol*) je číslo jednoznačně identifikující síťové rozhraní počítače. Může se jednat přímo o adresu konkrétního počítače nebo skupinu počítačů přistupující do sítě přes jednotný uzel.

^d Ve většině případů se jedná o restart serveru.

6.5.4 Technologický popis

Za účelem využití programu analýzy více pracovníky bylo pro přístup vytvořeno lokální webové uživatelské rozhraní (obr. 6.5). To zobrazuje aktuální soubory pro případnou analýzu, archivní soubory z analýz předchozích a umožňuje aktualizovat seznam současných hesel uložených v databázi ze souboru.

Rozhraní je postaveno na programovacím frameworku Django propojeném s MySQL databází, resp. se dvěmi. MySQL databáze s daty se skládá ze dvou samostatných tabulek. První obsahuje aktuální výsledky, druhá termíny již analyzované a posouzené jako kandidáty na nová hesla. Druhá databáze je základní databází PSH, která slouží pro kontrolu shody vyhledávaného termínu s některým ze znění předmětového hesla PSH. Starší výsledky jsou uloženy v CSV souboru^a v konkrétním adresáři, stejně tak jako aktuální – ještě nezpracované logy.

^a CSV soubor (Comma Separated Values, *čárkou oddělené hodnoty*) je jedním ze základních formátů souborů pro ukládání dat. Obsahuje hodnoty oddělené čárkou (někdy nesprávně středníkem) s jednotlivými záznamy na samostatných řádcích (soused hodnot je stejný pro všechny řádky).

Analýza logu vyhledávání v katalogu NTK

Vložte CSV soubor se zpracovanými hesly (".csv")

Soubor nevybrán

Nezpracované logy

- [www_server_4991.log.1205.1527](#) (196 kB)
- [www_server_4991.log.1806.1142](#) (788680 kB)
- [www_server_4991.log.2608.1308](#) (7873 kB)
- [www_server_4991.log.1205.1520](#) (170284 kB)
- [www_server_4991.log.0511.1210](#) (534551 kB)
- [www_server_4991.log.0411.1209](#) (933934 kB)
- [www_server_4991.log.0805.1751](#) (6 kB)
- [www_server_4991.log.0610.1138](#) (740163 kB)
- [www_server_4991.log.0805.0842](#) (211094 kB)
- [www_server_4991.log.2508.1325](#) (1000837 kB)
- [www_server_4991.log.1911.1048](#) (351410 kB)

Archiv výsledků

- [logy-2011-03-30.csv](#) ×

Obr. 6.5 Webové rozhraní analýzy logu vyhledávání

Analýza logu vyhledávání v katalogu NTK

🌀 Právě probíhá analýza logů. Mějte strpení. Může to trvat pěkně dlouho...

Obr. 6.6 Probíhající analýza logů

Analýza logu vyhledávání v katalogu NTK

CSV soubor ke stažení: [logy-2011-03-30.csv](#)

Celkový počet termínů: 2883

[Zpět na hlavní stránku ->](#)

Obr. 6.7 Výsledek analýzy logů

6.5.5 Postup práce

V případě ukončení používání současného logu a jeho vytvoření, je soubor nahrán do určeného adresáře na serveru, který je kontrolován pro výskyt nezpracovaných souborů. Tyto soubory se zobrazují na hlavní stránce rozhraní a je možné spustit jejich analýzu. Po zpracování všech aktuálních logů je výsledek uložen do CSV souboru s počtem výskytu termínu a pevně stanoveném počtu oddělovačů pro doplňované hodnoty. Soubor je následně nabízen ke stažení spolu s informací o počtu nových termínů a celkovém počtu neposouzených termínů v databázi s aktuálními hesly (obr. 6.7). V případě jeho ztráty je jeho stažení nadále možné z oddělení archivních výsledků přímo z rozhraní. LOG soubory se z důvodu velikosti a omezeného množství kapacity pro jejich uložení po zpracování automaticky odstraňují.

K posuzování termínů dochází kolaborativně prostřednictvím služby Google Docs. Výsledný CSV soubor se do této služby importuje a určí se nejnižší prahová hodnota výskytu pro posuzování hesla (obr. 6.8). Při posuzování je ke každému z nich zapsán důvod jeho zařazení či nezařazení do PSH, v případě zařazení i jeho umístění a role (preferované, nepreferované).

Po zpracování celé dávky se vytvoří CSV export, který se nahrává přes webové rozhraní zpět do databáze. Zpracované termíny se vyřadí z tabulky pro aktuální hesla a zařadí se mezi již posouzené termíny. Ostatní termíny zůstávají nadále v tabulce se současnými hesly a jsou posuzovány, až když jejich multiplicita dosáhne prahové hodnoty. Některé ze sporných termínů jsou za účelem nového připomenutí také ponechány mezi aktuálními daty.

	počet	zkratka	termín	N/D	kontrol. č.	pozn.
	22		amatérské rádio			název časopisu
23.11.201	22	sr	alternativní pohony	D	13891	navrhuji buď jako podřízené k "pohony strojů"
	21		vzduchotechnika			už máme "větrání a vzduchotechnika"
	21		nauka o budovách			už máme "budovy"
	21		etics			produkt
	21		chip			název časopisu
	21		acta polytechnica			název časopisu
	20		venturi			osobnost, automobilka
	20		ubuntu			produkt
	20		statické tabulky			název dokumentu
	20		německy s úsměvem			název dokumentu
	20		mysql			produkt
	20		feynman			osobnost
	20		docbook			produkt
23.11.201	20		auto	N		navrhuji navrhuji přiřadit nedeskriptor "auta" k heslu "automobily"
	20		ajax			produkt
	19		čsn			zkratka pro normy
	19		pozemní komunikace			skripta, navíc už máme hesla "stavby pozemních komunikací", "silniční komunikace" a "veřejné komunikace"
	19		mobilní sítě			už máme nedeskriptor "mobilní sítě 3g"
	19		kevin lynch			osobnost
	19		kalina			osobnost
	19		geodezie			už máme deskriptor "geodézie"
	19		elektrické stroje a pohony			už máme deskriptory "elektrické stroje" a "elektrické pohony"
	19		eb			nevím, co se tím myslí. v opacu máme nějaké brusky; google našel virus, vyrážku a různé firmy...
	19		eisee			produkt, korporace
23.11.201	19		chemie potravin	N		navrhuji navrhuji přiřadit jako nedeskriptor k heslu "potravinářská chemie", i když může to být hledání stejnojmenných skript
	19		atmel avr			produkt
	18		koucký			osobnost
	18		koolhaas			osobnost
	18		business intelligence	?		navrhuji zařadit pod "podnikání", už existuje jako produkt
	18		armstrong			osobnost

Obr. 6.8 Sdílená tabulka pro vyhodnocování termínů

6.5.6 Algoritmus analýzy

Analýza logů se opírá o extrakci textu za využití regulárních výrazů^a. Soubor je zpracováván sekvenčně (po řádcích) a společně s nalezeným požadavkem do katalogu je ukládána také IP adresa uživatele a datum a čas požadavku.

Vyhledávací parametry jsou omezeny na vyhledávání ve všech polích, předmětu a heslu PSH. Automaticky jsou vyřazeny také termíny obsahující číslovky.

^a Regulární výraz (zkráceně *regexp*, z anglického *regular expression*) je uživatelsky definovaná šablona textu, která se používá k vyhledávání textových vzorů a modifikaci textu. Jedná se o univerzální nástroj, který je implementován ve většině programovacích jazyků, stejně jako i řadě textových editorů.

Při nalezení dotazu do katalogu se termín extrahuje a v případě, že je shodný s předchozím nalezeným termínem, zkontroluje se rozdílnost IP adres, popřípadě délka časového intervalu mezi požadavky. Tato kontrola probíhá z důvodu návratu uživatele ze zobrazeného záznamu na výsledek vyhledávání. V této chvíli totiž dochází k opětovnému odeslání požadavku. Jeden uživatel tak odešle stejný řetězec hned několikrát za sebou. Pokud se IP adresy neshodují nebo pokud je časový interval dostatečný, je řetězec přijat jako potenciální kandidát na nové heslo.

Před uložením do databáze je nalezený textový řetězec standardizován převedením na malá písmena a nahrazením zástupných znaků ve tvaru např. „%C3“ za skutečná písmena. Tyto znaky v textu vznikají z důvodu použití kódování, ve kterém nejsou definované české znaky^a. K překladu dochází za pomoci jednoduché tabulky (Python slovníku^b).

Před zařazením do databáze se současnými daty se nejprve kontroluje, zda termín již neexistuje mezi hesly PSH nebo zda již nebyl posuzován při předchozích analýzách. Pokud ano, je odstraněn. V opačném případě dojde k jeho uložení do databáze s aktuálními daty - v podobě nového termínu s multiplicitou jedna nebo jejím zvýšením o jedna u již přítomného.

^a Stejně jako všechna ostatní data je text v počítači uložen v binárním kódu (nuly a jedničky), tedy v číslech. Kódování textu by se dalo přirovnat k překladové tabulce, která určuje jaký znak přiřadit k danému číslu. Různá kódování se liší počtem bitů použitých pro jeden znak (v překladu: rozsahem hodnot, počtem definovaných znaků) nebo rozdílnou definicí znaku pro stejné číslo. V případě, že zobrazujeme text v jiném kódování, než v jakém byl uložen, program ohlásí chybu nebo část neznámých znaků nahradí jinými symboly. Samozřejmě se může stát i to, že je text zobrazen správně, a to v případě, že jsou znaky v těchto kódováních definovány stejně.

^b Slovník (*dictionary*) je v programovacím jazyce Python jednou za základních datových struktur. Je tvořen dvojicemi „hodnota“ : „klíč“. Textová reprezentace slovníku je prakticky shodná s datovým formátem JSON.

6.5.7 Výsledky analýzy

Následující výsledky jsou výstupem analýzy logů za přibližné časové období 3 měsíců. Za tuto dobu bylo vyextrahováno 1653 termínů, které odpovídali nastaveným parametrům (tabulka 6.2). Mezi nimi se logicky nacházela velká část na první pohled nepoužitelných hesel. Jednalo se převážně o vyhledávání názvů knih, jmen autorů, řada překlepů, názvy nakladatelů, geografická jména a další.

Filtrováním všech získaných termínů na konci zůstala přibližně třetina z nich (567), které byly posuzovány jako kandidáti na nová předmětová hesla. Velká část z nich se již v hesláři vyskytovala jinak vyjádřena, v jiné formě. Některé termíny byly pro naše potřeby příliš detailní, jiné popisovaly velmi málo zastoupená témata. Objevila se zde také řada spojení, která byla zamítnuta již dříve při analýze klíčových slov.

Celým procesem nakonec prošlo 13% - 74 z posuzovaných hesel. 12 z nich již bylo zařazeno jako deskriptor a 49 jako nedeskriptor mezi předmětové authority, 13 stále čeká na společné posouzení vhodnosti a místo zařazení.

Analýza potvrdila malé využívání rozšířeného hledání v knihovním katalogu a při vyhledávání na internetu obecně. Jako její vedlejší produkt však bylo navíc zakoupeno několik žádaných knih do knihovního fondu.

6.5.8 Současnost

V současné době máme k dispozici další soubory logů za období 8 měsíců o celkové velikosti 5 GB. Vzhledem k počtu extrahovaných hesel a nutné manuální práci trvá celý proces zařazení hesla poměrně dlouhou dobu, která je však s cílem udržení kvality a rozsahu hesláře v určitých mezích, nezbytná. Z tohoto důvodu se v budoucnosti patrně bude zvyšovat minimální hranice výskytu hesla pro zařazení mezi kandidáty.

Extrahované termínů s multiplicitou větší než 3		1653
Vyhledávání podle	předmětová hesla, klíčová slova	567
	název (knihy, skripta, periodika, články)	472
	osoba (od autora XY, o osobnosti XY)	338
	produkt	138
	jiné	110
	firma, ateliér, instituce, nakladatelství, konference, výstavy	22
	geografické jméno	6
Předmětová hesla	schválené deskriptory	12
	schválené nedeskriptory	49
	čekající	13
	zamítnutá	493
Předmětová hesla zamítnutá	už máme	303
	příliš podrobné	57
	vágní	27
	zamítnuté klíčové slovo	32
	máme o tématu málo dokumentů	24
	jiné	50
Název	knihy	222
	skripta	161
	periodika	87
	články	2

Tabulka 6.2 Výsledky analýzy logu vyhledávání

7 Presentace PSH

Struktura a použití systému nabízí mnoho možností pro své grafické zobrazení. Jediným z dosavadních je Prohlížení PSH (viz. kapitola 3.7), které zobrazuje grafy jednotlivých hesel na základě jejich provázanosti. Katalogizačním požadavkům však tato reprezentace nevyhovuje kvůli vytržení konceptu z širší přilehlé struktury a špatné orientaci v zobrazení.

Prohlížení PSH představuje pouze jeden z aspektů pohledu na systém. Především nezachycuje rozdíl mezi využitím jednotlivých hesel v rámci katalogu, který by se dal nazvat váhou hesla, tématu.

Vzhledem k prostorovým nárokům současné prezentace hesel v Prohlížení hesláře bylo rozhodnuto o jeho přesunu na samostatnou stránku, která usnadní celkovou orientaci v rozhraní. Zároveň také zůstane místem pro odkazování z vnějších systémů. Na původním místě vznikne nová grafické zobrazení, vyhovující podmínkám a efektně zobrazující tematické zaměření fondu NTK.

7.1 Budoucnost Prohlížení PSH

Mimo přesunu na samostatnou stránku projde Prohlížení PSH řadou kosmetických a funkčních změn. Změny ve vzhledu se týkají několika úprav v rozmístění prvků na stránce. Patrně největší z nich bude přemístění výsledku vyhledávání dokumentů v katalozích NTK a NK těsně pod grafické zobrazení. V současnosti se totiž často z důvodu většího počtu přidružených hesel stává, že tento seznam byl mimo viditelnou část plochy. Uživatel tak neví, že Prohlížení PSH tuto funkci nabízí. Dále budou přidány odkazy na hesla ve Wikipedii. Provázání bylo docíleno částečně z již vytvořených vazeb na DBPedit a částečně mapováním na česká hesla. V případě použití vazeb na DBPedit bylo zjišťováno, zda je konkrétní článek k dispozici i v českém jazyce.

K funkčním změnám dojde především ve způsobu vyhledávání. Nově bude implementován našeptávač hesel PSH (po technologické stránce stejný jako v případě PSH Manageru (využití Django Frameworku a jQuery, viz. kapitola 7.3.1), jen jazyk vyhledávání se liší podle aktuálně nastavené jazykové mutace).

U výsledků vyhledávání nově přibude v závorce uvedený nedeskriptor pro shodu podle nepreferovaného znění.

7.2 Nová prezentace PSH na stránkách NTK

Grafická prezentace hesláře na stránkách NTK je limitována omezeným místem v základním zobrazení NTK webu. To je dáno navigačními menu, postranními odkazy a především fixní šířkou šablony.

Grafická prezentace jednotlivých úrovní PSH byla vybírána na základě požadavku na odlišení váhy hesel podle počtu výskytů v bibliografických záznamech. Z možností, které se nabízely v podobě různě tvarovaných sloupcových grafů, bublinových grafů a treemap, nakonec byl vybrán tag cloud. Tag cloud je na webu velmi oblíbená grafická prezentace zachycující váhu termínů v textovém seznamu.^{a 63} K tomu může docházet na základě velikosti písma, intenzity a typu barvy, nebo jejich spojením. Využívané termíny jsou velké a sytě barevné a nepoužívané naopak malé a průsvitné.

7.2.1 Technologický popis

K vytvoření tag cloudu (obr. 7.1) byl použitý již vytvořený JavaScript plugin – jQCloud^b. jQCloud implementuje prezentaci tag cloudu jen s použitím HTML a CSS. Z tohoto důvodu odpadá řešení nekompatibility v různých prohlížečích a na různých platformách. CSS také umožňuje bez problémů měnit jeho vzhled (barva a velikost písma pro různé váhy hesel).

Na svém vstupu přijímá algoritmus seznam hesel ve formátu JSON. Pro každé heslo je potřeba zadat tři atributy: textová reprezentace hesla, váha hesla a URL, v případě, že heslo odkazuje na další stránku. Váhy nejsou číselně omezeny. Program je sám přepočítává a rozděluje hesla do deseti úrovní. Nejvyšší úroveň znamená velká písmena a syté barvy, nejnižší naopak. Vše je však možné uživatelsky nastavit i jen s částečnou znalostí kaskádových stylů.

^a Jeho nejčastější využití bývá k zobrazení zastoupení klíčových slov nebo počtu vyhledávání pro jednotlivé termíny.

^b <https://github.com/DukeLeNoir/jQCloud>

7.2.2 Současný stav

V současnosti (1.4.2011) je nové prohlížení PSH na stránkách NTK částečně připraveno ke spuštění, musí však být dořešeny některé ze základních implementačních postupů. To hlavně z hlediska celkového obsahu stránky a údajích o heslu. Vzhledem k požadavkům na jednoduchost a „čistotu“ zobrazení patrně budou hesla ochuzena o linky na vnější systémy, nepreferovaná znění a možná i hesla příbuzná. Tag cloud by měl tedy v rozhraní zůstat samotný jen s odkazem na nadřazené heslo a výsledkem hledání v katalogu NTK.



Obr. 7.1 Ukázka váženého tag cloudu pro nejvyšší úroveň PSH

7.3 PSH Manager

PSH Manager vznikl jako nástroj určený pro katalogizátory a další aktivní uživatele hesláře. Byl vytvořen podle specifických požadavků na vzhled a funkčnost. Popud k jeho vzniku dali sami katalogizátoři, když vyjádřili nespokojenost se současným prohlížením hesláře. Zároveň odkazovali na původní verzi programu PSH Manager, která jim funkčně vyhovovala (tato skutečnost jistě pramenila i ze zvyku). PSH Manager byl tedy definován jasnými požadavky na vzhled i implementované funkce.

7.3.1 Technologický popis

Technologické prostředky určené k vytvoření PSH Manageru vycházeli z jeho charakteristiky jako webové aplikace. Online přístupem k hesláři odpadá problém s jeho distribucí ostatním institucím, stejně tak jako nutnost instalace programových nástrojů a knihoven na cílových počítačích uživatelů.^a

PSH Manager je postaven na programovacím jazyce Python, resp. programovacím frameworku Django, který se stará o komunikaci s databází a dynamické vytváření webových stránek. Celý projekt běží na linuxovém serveru s nejrozšířenějším HTTP (*webovým*) serverem Apache s modulem `mod_python`, který je určený pro práci s daným programovacím jazykem. Většina operací je tedy prováděna na serveru (*server-side*^b).⁶⁴

Heslář je uložený v relační databázi MySQL s automaticky vytvořeným modelem pro Django. To znamená, že selekce záznamů neprobíhá v podobě SQL příkazů, ale

^a V dnešní době již patrně nelze přemýšlet nad aplikacemi jinak než online. Aplikace používané skrze webový prohlížeč jsou již prakticky zavedeným standardem. Ulehčují správu, distribuci a okamžitý přístup prakticky odkudkoli s internetovým připojením (které je dnes doslova všude).

^b Server-side scripting (*skriptování na straně serveru*) je schopnost provádět výpočetní operace přímo na serveru, narozdíl od klient-side scriptingu (*skriptování na straně klienta*), kde jsou výpočty prováděny ve webovém prohlížeči, nejčastěji v JavaScriptu. Server-side metoda je ve většině případů rychlejší a bezpečnější.

přes Django Model Query API^a (*dotazovací API*) pro filtrování záznamů prostřednictvím databázového modelu. Primárním klíčem propojení tabulek jsou unikátní identifikátory hesel PSH (*číslo hesla s předponou PSH*). K aktualizaci databáze dochází automaticky pravidelně spouštěným skriptem. Nejprve je zjištěno nejvyšší číslo hesla nacházející se v lokální databázi. Následně se dohledá nejvyšší číslo hesla přístupné na serveru. V případě, že je toto číslo větší, nová hesla se stáhnou a doplní do lokální databáze i s dodatečnými vazbami pro již přítomné deskriptory. Jedná se o stejnou databázi, která je používána jako zdroj dat pro prohlížení hesláře na stránkách NTK.

Výsledné webové stránky jsou napsány ve značkovacím jazyce HTML5 s kaskádovými styly CSS3, formátujícími vzhled. Pro komunikaci se serverem a grafické efekty je využitý JavaScript framework jQuery. jQuery komunikuje se serverem synchronně i asynchronně (*Ajax*) podle konkrétní funkce.^b Vzhledem k implementaci není možné stránky procházet bez zapnutého JavaScriptu v prohlížeči.

7.3.2 Uživatelské rozhraní

Uživatelské rozhraní manageru (obr 7.2) je koncipováno se zaměřením na obsah. To znamená, že se snaží zobrazit co největší část ze struktury hesláře. Kontext hesel je při klasifikaci dokumentů velmi důležitý. Uživatelskému rozhraní dominují tři hlavní prvky: nástrojová lišta, navigační strom hesláře a reprezentace konceptu hesla.

^a API (Application Programming Interface, aplikační programovací rozhraní) je sbírka funkcí („*příkazů*“) určená pro používání konkrétního programovacího nástroje.

^b Asynchronní komunikace se serverem je již delší dobu velmi oblíbeným, a hlavně efektním a efektivním způsobem vytváření HTML stránek. Hlavní rozdíl oproti klasické synchronní komunikaci je možnost změny obsahu stránky bez nutnosti jejího opětovného načtení. Vzhledem k tomu, že se ve většině případů velká část stránky nemění, můžeme tímto způsobem rapidně snížit objem dat vyměňovaných mezi server a klientem. Tento způsob navíc umožňuje provádět operace v pozadí, zatímco uživatel může pracovat s již načteným obsahem. Mezi hlavní nevýhody patří zvýšení počtu HTTP požadavků mezi klientem a serverem, omezení ve zpětné navigaci stránkami a odkazování na stránky vytvořené AJAXem, které není nemožné, ale mnohem složitěji implementovatelné.⁶⁵

7.3.2.1 Navigační lišta

V navigační liště se nachází funkční tlačítka, aktuální odkazy na tři poslední zobrazená hesla, odkazy na stránky PSH a NTK, a vyhledávací pole s integrovaným našeptávačem. První z tlačítek (*ikona knihy*) slouží pro zobrazení dokumentů v katalogu NTK, které mají přiřazené aktuálně zobrazené heslo. Druhé tlačítko (*ikona se znakem plus*) odkazuje na formulář pro návrh nového hesla pro případ, že by uživatel chtěl dát podnět k zařazení chybějícího termínu do hesláře. Historie zobrazených hesel slouží k jednoduchému návratu na tři poslední, aniž by je uživatel musel vyhledávat ve stromové struktuře nebo zadávat do vyhledávacího pole.

Vyhledávací pole s integrovaným našeptávačem^a nabízí preferované a nepreferované formy hesel v češtině nebo v angličtině (obr. 7.3, 7.4). Vždy je vyhledáváno pouze v jednom jazyce. K jeho nastavení slouží tlačítko vedle vyhledávacího pole.

První v pořadí jsou termíny, které začínají na vyhledávaný řetězec, seřazené podle abecedy. Za nimi jsou hesla, která řetězec obsahují, ale nezačínají na něj (taktéž abecedně seřazena). Zobrazeno je vždy jen dvacet prvních termínů tak, aby nepřesahovaly rámec zobrazené části stránky. Vzhledem k zobrazení možných hesel v průběhu hledání musí být požadavek na zobrazení celého konceptu absolutně shodný s deskriptorem či nedeskriptorem. V opačném případě heslo není nalezeno v databázi a dotaz zůstává bez odpovědi.

K vytvoření našeptávače byl využitý již vytvořený plugin jQuery Autocomplete. Autocomplete se připojí k vyhledávacímu textovému poli s určením události, která jej vyvolává. V našem případě „*keyup event*“ („puštění klávesy“). Po zaslání požadavku na server, se vyhledají požadovaná hesla, která se vrací v podobě

^a Našeptávač je funkce textových polí, která v reálném čase nabízí možné výsledky hledání obsahující aktuální řetězec např. podle odpovídajících položek uložených v databázi. Tato funkce je umožněna již dříve zmíněným asynchronním dotazováním na server, které dovoluje měnit aktuální obsah stránky bez jejího znovunačtení.

seznamu ve formátu JSON^a. Plugin se již o zbytek věcí postará sám. Navíc můžeme specifikovat jeho další atributy jako například minimální počet znaků, který vyvolá vyhledávání (nastaveno na tři znaky). K výslednému požadavku na zobrazení reprezentace hesla dochází až při stisknutí klávesy enter.

7.3.2.2 Navigační strom

Navigační strom je tvořen celou hierarchickou strukturou hesláře. V základním stavu jsou vidět pouze hlavní tematická hesla. Po kliknutí na jednotlivé heslo se rozbalí konkrétní část podstromu v případě, že jsou na konkrétní heslo navázána hesla podřazená. Tato skutečnost je indikována modrou šipkou vlevo od termínu. Při kliknutí dochází ke dvěma událostem. Zaprvé ke zobrazení kompletní reprezentace konceptu a zadruhé k posunu stromu („scrollingu“), tak aby bylo aktuální (zvýrazněné) heslo přibližně na středu obrazovky. Tato událost nastává při všech příležitostech výběru. To znamená při zadání přes vyhledávací pole, při výběru přes historii hesel a při kliknutí na odkaz v reprezentaci hesla.

Navigační strom není tvořen dynamicky (jak by se mohlo nabízet při velkém počtu hesel a postupném zobrazování struktury), ale je statický. Je uložený v souboru a při otevření stránky se pouze vkládá na své místo. Tato metoda byla zvolena na základě potřeby rozbalování stromu na aktuálním termínu, a to včetně zobrazení všech rodičovských hesel spolu s dalšími deskriptory na stejné úrovni.

Vzhledem k možné výšce stránky při násobném rozbalení podstruktur v aktuálním čase je navigační strom oddělený ve své vlastní „scrollovatelné“ části. O výšce stránky rozhoduje jen zobrazený koncept, který ve většině případů nepřesáhne rozměry monitoru.

7.3.2.3 Reprezentace konceptu

Poslední část uživatelského rozhraní představuje textová reprezentace konceptu. Jedná se o prostou tabulku s výpisem všech vazeb aktuálního hesla. Preferovaná a

^a JSON (JavaScript Object Notation, *JavaScriptový objektový zápis*) je jednoduchý textový formát pro výměnu dat. Jak název napovídá, vychází ze zápisu objektu v programovacím jazyce JavaScript (přesto používaný globálně). Je velmi dobrou alternativou k zápisu XML, protože je nejen dobře čitelný, ale také textově úspornější.

nepreferovaná znění jsou vypsána v českém a anglickém jazyce, nadřazené, podřazená a příbuzná hesla pouze česky. Hierarchická a příbuzensky vázaná hesla jsou aktivní a je možné je otevřít přímo z tabulky.

Popis hesla je z důvodu omezení počtu dotazů do databáze a zrychlení rozhraní načítán ze souboru (staticky). Všechna hesla jsou uložena v adresáři na serveru, identifikována svým PSH id (např. *PSH1562.html*). Soubor obsahuje část HTML kódu, který je vložen na své místo do výsledné stránky.

K vytvoření reprezentace v reálném čase (dynamicky, asynchronně) dochází v případě, že program nenalezne soubor s daným číslem. To znamená, že se po kliknutí na otevření konceptu, odešle dotaz na server (*HTTP request^a*), který zjistí, že statická reprezentace pro konkrétní heslo neexistuje. Následuje dotaz do databáze, zjištění potřebných vazeb, vytvoření reprezentační tabulky a její vracení v podobě HTTP odpovědi (*HTTP response^b*).⁶⁶

Vzhledem k implementaci rozhraní není možné odkazovat na aktuální stránku – aktuální heslo. Veškerý obsah je vkládán do současné stránky a její adresa se nemění. Při odkázání se z tohoto důvodu vždy nacházíme na úvodní obrazovce. Nicméně je na hesla možné odkazovat jiným způsobem, a to pomocí adresy ve tvaru *data.ntkcz.cz/psh_manager_online/concept/PSHXXXX*. Při zavolání URL v tomto tvaru se do strany vloží heslo označené požadovaným číslem PSH.

^a HTTP požadavek (HTTP request) je dotaz na server s definovanou strukturou. Skládá se z určení použité metody, požadované adresy, verze protokolu http a adresy serveru. Další informace jako např. kódování, jazyk nebo formát dat jsou v nových verzích HTTP volitelné.

^b HTTP odpověď (HTTP response) je odpověď serveru na HTTP požadavek. První tzv. stavový řádek specifikuje verzi HTTP, číselné vyjádření odpovědi serveru (200 – stránka nalezena, 404 – stránka nenalezena) a textovou reprezentaci stavu (200 OK, 404 NOT FOUND). Další řádky slouží pro dodatečné informace o serveru a o datech. Po jedné prázdné řádce následují samotná data (text, soubor).

- ❖ astronomie
- ❖ biologie
- ❖ chemie
 - ❖ agrochemie
 - ❖ analytická chemie
 - ❖ anorganická chemie
 - ❖ biochemie
 - ❖ chemický průmysl
 - ❖ ekologická chemie
 - ❖ farmaceutická chemie
 - ❖ fyzikální chemie
 - ❖ jaderná chemie
 - ❖ kvantová chemie
 - ❖ makromolekulární chemie
 - ❖ obecná chemie
 - ❖ **organická chemie**
 - ❖ deriváty uhlovodíků
 - ❖ funkční skupiny
 - ❖ organická syntéza
 - ❖ přírodní látky
 - ❖ supramolekulární chemie
 - ❖ uhlovodíky
 - ❖ petrochemie
 - ❖ potravinářská chemie
 - ❖ toxikologie
 - ❖ dmonrava

organická chemie

organic chemistry

Nadřazené heslo

chemie

Nepreferované znění [CS]

-

Nepreferované znění [EN]

-

Příbuzná hesla

biochemie
chemické sloučeniny
farmaceutická chemie
potravinářská chemie

Podřazená hesla

deriváty uhlovodíků
funkční skupiny
organická syntéza
přírodní látky
supramolekulární chemie
uhlovodíky

Obr. 7.2 Webové rozhraní PSH Manageru

chemie
 chemie a chemická technologie
 chemie potravin
 agrochemie
 analytická chemie
 analytická chemie životního prostředí
 anorganická chemie
 anorganická geochemie
 biochemie
 bioelektrochemie
 biofyzikální chemie
 biogeochemie
 ekologická chemie
 elektrochemie
 farmaceutická chemie
 farmakochemie
 fotochemie
 fytochemie
 fyzikální chemie
 geochemie
 geochemie Země
 geochemie endogenních procesů
 geochemie exogenních procesů
 geochemie izotopů
 geochemie minoritních prvků

- ❖ jaderná chemie
- ❖ kvantová chemie
- ❖ kvantová statistika
- ❖ makromolekulární chemie

fyzikální chemie

physical chemistry

Nadřazené heslo

chemie

Nepreferované znění [CS]

-

Nepreferované znění [EN]

-

Příbuzná hesla

chemická fyzika
 struktura pevných látek
 termodynamika

Podřazená hesla

chemická termodynamika
 chemická vazba
 disperze
 dvojvrstva
 elektrochemie
 elektroosmóza
 fotochemie
 hydrochemie
 mechanochemie
 rozpouštění
 struktura hmoty
 ultrafiltrace

Obr. 7.3 Našeptávač - vyhledávání v českém jazyce

chemistry
chemistry
chemistry and chemical technology
Earth geochemistry
agrochemistry
analytical chemistry
biochemistry
bioelectrochemistry
biogeochemistry
biophysical chemistry
carbon (chemistry)
carbonization (chemistry)
centrifugation (chemistry)
colloid chemistry
columns (chemistry)
constitution (chemistry)
cosmochemistry
derivatives (chemistry)
drugs (chemistry)
electrochemistry
environmental chemistry
food biochemistry
food chemistry
fuels (chemistry)
general chemistry

- ultrafiltrace
- ❖ jaderná chemie
 - ❖ kvantová chemie
 - ❖ kvantová statistika
 - ❖ makromolekulární chemie

fyzikální chemie *physical chemistry*

Nadřazené heslo

chemie

Nepreferované znění [CS]

-

Nepreferované znění [EN]

-

Příbuzná hesla

chemická fyzika
struktura pevných látek
termodynamika

Podřazená hesla

chemická termodynamika
chemická vazba
disperze
dvojvrstva
elektrochemie
elektroosmóza
fotochemie
hydrochemie
mechanochemie
rozpouštění
struktura hmoty
ultrafiltrace

Obr. 7.4 Našeptávač - vyhledávání v anglickém jazyce

8 Nové možnosti využití hesláře

Při zkoumání nových možností využití hesláře se nabízela možnost na spolupráci s projektem Národního úložiště šedé literatury (NUŠL). Tato spolupráce spočívala ve vlastní implementaci automatického indexeru, který by po analýze plného textu přiřadil dokumentu nejvhodnější hesla PSH.

8.1 Automatická indexace dokumentů v NUŠL

Web umožnil autorům šedé literatury^a velmi rychle a levně publikovat svá díla. S touto možností by však měla přicházet i odpovědnost svá díla označit, aby mohla být dobře vyhledána.⁶⁷ To je samozřejmě v zájmu autorů, kteří chtějí, aby jejich články byly čteny, stejně jako v zájmu čtenářů, kteří chtějí na svůj dotaz dostat relevantní informace. K tomuto procesu však většinou nedochází a jejich články bývají indexovány špatně nebo vůbec. Dokumenty spadající pod šedou literaturu jsou méně viditelné, neboť nejsou distribuovány běžnými mechanismy, jako jsou komerční databáze aj. Předmětová indexace by měla umožnit lepší dohledatelnost těchto dokumentů přes knihovní katalogy i běžné internetové vyhledávače.^{68 - 70}

Je obecně známo, že nejlepším způsobem klasifikace dokumentů je využití souboru předmětových hesel. Autoři znají dobře obsah svých děl, s obsahem konkrétního znalostního systému často však seznámeni nejsou nebo ho nepoužívají vůbec. Velmi důležitá je jistě i znalost kolekce dokumentů, mezi které autor své dílo zařazuje.

V dnešní době však již nemusí být klasifikace dokumentů nutně zdlouhavým a nudným procesem. Existuje řada softwarových nástrojů, které pomáhají tento

^a „Šedá literatura představuje rozmanitou kolekci dokumentů publikovaných na všech úrovních vládní, akademické, firemní a průmyslové sféry, v papírové i elektronické formě, které jsou chráněny právy intelektuálního vlastnictví, dostatečně kvalitní na to, aby byly sdružovány a uchovávány v knihovních fondech nebo institucionálních repozitářích, avšak ne kontrolovány komerčními vydavateli.“ [GL12, 2010] Primárně se tedy jedná o bakalářské, diplomové a dizertační práce, sborníky z konferencí, výroční zprávy a další. Je velmi úzce spojena s volným přístupem k dokumentům na internetu pod názvem „open access“ (*volný přístup*).

proces urychlit a automatizovat. Ty navíc mohou dosahovat stejně dobrých indexačních výsledků jako proces manuální klasifikace. Většina těchto nástrojů je optimalizována pro práci s anglickým jazykem (což je logické), ale nabízejí možnosti k jejich přizpůsobení většině světových jazyků, při poskytnutí základních funkčních částí pro konkrétní jazyk.

8.1.1 Automatická indexace

Cílem automatické indexace je zjednodušení celého procesu na základě odstranění požadavků na znalost organizačního systému a dokumentů v kolekci. Výsledkem by měl být záznam zařazený do sbírky kvalitně a konzistentně. Neměl by však autorovi brát právo autonomního určení přiřazených předmětových deskriptorů. Autor má nad hesly plnou kontrolu a i samotný proces automatické indexace je pouze volitelný – nepovinný.

Přístup, který v NTK používáme je „kontrolované strojové učení“ (*supervised machine learning*). Indexační model vychází z kolekce již indexovaných dokumentů podle použitého systému. Na základě této trénovací množiny^a se software pro automatickou indexaci učí, jak „správně“ indexovat. Esenciální je v tomto případě přístup k plnému textu dokumentu. V běžné praxi by se jednalo o problém. V námi navrhovaném konceptu, však indexuje dokument sám autor. Práva k němu tak jsou automaticky udělena.

8.1.2 Technologický popis

Automatická indexace se skládá z několikastupňové sekvence úkonů, které provádějí základní softwarové části mechanismu. Na jejím počátku je soubor s plným textem a na jejím konci seznam přiřazených předmětových hesel. Používány jsou samostatné softwarové komponenty, které jsou propojeny dodatečným programovým kódem. V některých případech bylo nutné zasáhnout i

^a Trénovací množina je obecně množina prvků určených pro tvorbu datově-třídícího modelu při kontrolovaném strojovém učení. Podle něj jsou následně tříděny (kategorizovány) prvky stejného druhu.

do samotných komponent. To by nebylo možné bez použití *open source*^a nástrojů s modulární architekturou^b. K programovému vybavení musíme připočítat také datové zdroje v podobě předmětových hesel a kolekce dokumentů pro stavbu indexačního modelu.

8.1.2.1 Systém předmětových hesel

Systémem předmětových hesel je PSH (viz. kapitola 3) ve formátu SKOS. Za účelem indexace byl heslář upraven, konkrétně byla odstraněna tematická větev „obecnosti“, která by v důsledku svého obecného, nepříliš specializovaného zaměření podporovala falešně pozitivní výběr deskriptorů.

8.1.2.2 Digitální repozitář

Digitální repozitář je hostitelským prostředím, které se stará o sbírku dokumentů. Je v něm také implementován celý proces automatické indexace, respektive jeho řízení, provázání komponent a rozhraní pro interakci s uživatelem.⁷¹

Námi používaný repozitář je software CDS Invenio^c,⁷² napsaný v Pythonu. Právě díky jeho modulární architektuře jej bylo možné doplnit o potřebné funkce. Invenio propojuje všechny jednotlivé části, stará se o manipulaci s dokumentem a vytváří uživatelské rozhraní služby.

8.1.2.3 Automatický indexer

Automatický indexer^d je hlavní částí zodpovědnou za klasifikaci dokumentu. V našem případě se jedná o software Maui Indexer. Jeho autorka, Olena Medelyan, tvrdí, že Maui dokáže generovat výsledky schopné konkurence k manuální

^a Open source (*otevřený kód*) označuje software s přístupným zdrojovým kódem. Umožňuje uživateli za určitých podmínek kód prohlížet a upravovat.

^b Modulární architektura (modular architecture) označuje programy skládající se z jednotlivých samostatných částí. Moduly je možné doplňovat i odstraňovat, pokud na nich některé z dalších nejsou funkčně závislé.

^c <http://invenio-software.org/>

^d <http://code.google.com/p/maui-indexer/>

indexaci, které dokládá jejich srovnáním s ručně přiřazenými hesly.^{a 73, 74} Je napsaný v programovacím jazyce Java.

Maui Indexer je prezentován jako nezávislý na použitém jazyku. Avšak pro co nejlepší výsledky je nutné integrovat jazykově specifické komponenty. Vzhledem k poli působnosti byl klasifikátor adaptován na český jazyk. Dvě hlavní jazykově specifické části jsou: množina *stopwords*^b a český *stemmer*^c.

Množina *stopwords* byla získána extrakcí slov z Českého národního korpusu^d, konkrétně ze 700 000 000 publicistických textů v rozmezí let 1995 – 2007 (korpus *SYN2009PUB*). Tato slova byla následně ručně revidována. Český národní korpus představuje obrovskou sbírku dokumentů odrážející současný český psaný projev.⁷⁵ V ideálním případě se však množina *stopwords* získává přímo z korpusu, s nímž program pracuje.

Za účelem redukce textu a sdružování slov se stejným kořenem bylo nutné adaptovat *stemmer* pro český jazyk. Adaptován byl agresivní český *stemmer*^{e 76} (*agressive czech stemmer*), který bohužel podle prvních testů nefunguje plně podle předpokladů. Při dalším testování bude zváženo, zda se tento *stemmer* bude při indexaci používat nebo se pokusíme o integraci jiného.

8.1.2.4 Kolekce textů

K vytvoření indexačního modelu musíme získat „dostatečný“ počet plných textů článků označených hesly PSH. Taková množina bohužel není lehce obstaratelná a všechny články musí být získávány ručně na základě katalogu NTK. V něm byly vyhledány záznamy článků označených hesly PSH, ke kterým byl následně

^a <http://code.google.com/p/maui-indexer/wiki/Examples>

^b Stopwords (*vyloučená slova*) je množina slov, které v kontextu daného dokumentu nemají význam. Tato slova jsou z dokumentu hned na začátku odebrána. Následná analýza probíhá na filtrované části textu. Nejčastěji se jedná o spojky, příslovce a pomocná slovesa.

^c Stemming (od anglického slova *stem - kmen*) je proces redukování slov na jejich kořen. Nemusí se nutně jednat o morfologický základ slova. Stačí, když se slova se stejným kořenem redukuji na stejný tvar.

^d <http://www.korpus.cz/index.php>

^e <http://members.unine.ch/jacques.savoy/clef/CzechStemmerAgressive.txt>

dohledáván text na internetu. V současné době je indexační model tvořen přibližně 50 texty.

8.1.2.5 Národní úložiště šedé literatury

NUŠL^a je digitální repozitář pro dlouhodobou archivaci a zpřístupnění šedé literatury. NUŠL je podporovaný Ministerstvem kultury v rámci projektu **"Digitální knihovna pro šedou literaturu - funkční model a pilotní realizace"**^b spravovaný NTK ve spolupráci s Vysokou školou ekonomickou. „Cílem NUŠL je nabídnout svým uživatelům odborné dokumenty šedé literatury s originálním a hodnotným obsahem.“ [FURSTOVÁ, 2010]

Repozitář obsahuje metadata a dokumenty ze sítě spolupracujících institucí v čele s Akademii věd a vysokými školami (Vysoká škola chemicko-technologická, Vysoká škola ekonomická a další). Implementace automatického indexeru v češtině je podmíněna primárně českým obsahem úložiště. Na základě spolupracujících institucí v současné době jeho obsah tvoří převážně bakalářské, diplomové, disertační práce a příspěvky z konferencí.⁷⁷

8.1.2.6 Uživatelské rozhraní

Uživatelské rozhraní je integrováno do digitálního repozitáře CDS Invenio. O zobrazení se stará speciálně navržená šablona zakomponovaná do standardního systému šablon Invenia. Funkčnost na uživatelské straně je částečně implementována v samotném repozitáři, částečně v Django frameworku (spravující našeptávač hesel), běžícím na webovém serveru Apache, a JavaScriptu s frameworkem jQuery, který spravuje manipulaci s HTML kódem na straně uživatele.

8.1.3 Proces indexace

Proces indexace iniciuje uživatel vkládající PDF dokument do repozitáře. V současné době je požadován plný text, který by měl poskytovat lepší možnosti pro zachycení a extrakci klíčových slov.^{78, 79} Ve vstupním formuláři zvolí, zda chce

^a <http://nrgl.techlib.cz>

^b <http://www.isvav.cz/projectDetail.do?rowId=DC08P02OUK007>

dokument automaticky indexovat (běžně nastaveno na zapnuto). V případě že ano, je PDF soubor převeden na čistý textový soubor a uložen do adresáře, který je určený pro dokumenty k indexaci. Následně je spuštěn Maui Indexer s nastavenými cestami k potřebným adresářům, indexačnímu modelu a systému předmětových hesel s definicí jeho formátu. Při indexaci je u dokumentu vytvořen nový soubor s navrhovanými předmětovými hesly.

8.1.4 Uživatelské rozhraní

V uživatelském rozhraní (obr. 8.1) se zobrazují navržená hesla s možností jakékoli z nich odebrat nebo přidat. K přidávání slouží našeptávač, který při psaní nabízí preferovaná i nepreferovaná znění hesel v češtině (obr. 8.2). Nepreferovaná jsou po vybrání přeložena na příslušný deskriptor a zobrazena v závorce vedle něj. Při vybrání hesla se ve spodní části stránky simultánně zobrazují dokumenty daným heslem již označené. Tato funkce by měla uživateli pomoci při určování vhodných deskriptorů na základě kontextu aktuální sbírky.

8.1.5 Výsledky testování

Výsledky testování se současným modelem mohou jevit poměrně rozporuplně. Při nízkém počtu přiřazovaných hesel vykazují správný indexační směr, avšak prozatím jen v menším počtu případů vybírají přesnější termíny na úkor obecnějších (např. „automatizace“ upřednostněno před „automatizace knihoven“). Při vyšším zvoleném počtu hesel, které chceme přiřadit, může dojít i k velmi nepřesným výsledkům. Hlavní příčinou je model, vytvořený z malého počtu dokumentů, který by se však časem měl postupně vylepšovat a zpřesňovat. Velkou roli hraje jistě i aktuální stemmer, který svými výsledky nepřesvědčil, a dochází tedy k implementaci a testování jeho náhrady.

8.1.6 Budoucí možnosti

V blízké budoucnosti započne práce na automatickém vylepšování indexačního modelu na základě dokumentů, kterým byla hesla při vkládání do úložiště přidělena ručně. Dále budeme zvažovat možnosti indexace plných textů nahrávaných automaticky z přidružených institucí. Je potřeba zvážit, jakým způsobem a podle jakého kritéria přiřazovat deskriptory, aby i bez lidské kontroly dosahovaly co nejvyšší kvality. V opačném případě by byla indexace vysoce kontraproduktivní. Zpětná vazba by měla sloužit i pro aktualizaci a vylepšování samotného hesláře.

Submit New Record

Submit New Record

1

end of action

SUMMARY(2)

Konferenční materiály

PSH Search

Přirazená hesla

- informační věda ✗
- automatizace knihoven ✗
- software ✗
- algoritmizace ✗

Dokumenty označené heslem 'informační věda' (celkem 65 dokumentů):

Polythematic Structured Subject Heading System

Do knihovny skrze webový prohlížeč

Using Environmental Grey Literature to Engage Public Participation in Decision Making: Government Transparency in Coastal Policy Development

Projekt NUSL a další projekty v ČR

Integration of an Automatic Indexing System within the Document Flow of a Grey Literature Repository

Portály v oboru Informační a knihovní vědy

Current Status and Issues in Collection and Dissemination of National Technical Reports in Korea: Improvement through the use of the National Science and Technology Information Service (NTIS)

From OpenSIGLE to OpenGrey : Changes and Continuity

Digitální knihovny - včera, dnes a zítra

Polytematický strukturovaný heslář

Finish

Submission no²: 1303298599_4013

Obr. 8.1 Uživatelské rozhraní automatické indexace

Submit New Record

Submit New Record

1

end of action

SUMMARY(2)

inform

informační systémy pro cestující

informace

informační věda

informační prameny

informační fondy

informační procesy

informační služby

informační potřeby

informační etika

informační systémy

informační síť

informační pracoviště

informační střediska

informatizace

informatika

informatické vzdělání

informování

informační společnost

informační management

informační zdroje

informační chování

informační systémy mezinárodní

informační systémy oborové

informační míry

informační technologie

informační bezpečnost

informační věda' (celkem 65 dokumentu):

system

gagne Public Participation in Decision Making: Government Transparency in Coastal Policy Development

em within the Document Flow of a Grey Literature Repository

Dissemination of National Technical Reports in Korea: Improvement through the use of the National Science and Technology Information Service (NTIS)

and Continuity

Finish

Submission no²: 1303298599_4013

Obr. 8.2 Rozhraní automatické indexace – našeptávač hesel PSH

71

9 Závěr

Polytematický strukturovaný heslář v současné době obsahuje 13 856 hesel. Za poslední rok bylo přidáno nebo upraveno skoro 500 z nich. Nově začali heslář používat Knihovna bankovního institutu a Univerzitní knihovna Pardubice.

K aktualizacím hesláře dochází průběžně celý rok, zatímco nová verze vychází vždy na jeho začátku. Byly automatizovány základní operace, které usnadňují práci při výběru hesel. Jednou z nich je vznik portálu pro analýzu logů vyhledávání v katalogu NTK. Ten umožňuje extrahovat vyhledávané dotazy dle stanovených parametrů, a porovnávat je s PSH a již dříve zpracovanými termíny. Výsledkem procesu je statistika, která slouží jako podklad pro výběr kandidátů na nová hesla. Tento zdroj se osvědčil také při doplňování knihovního fondu NTK.

Mezi další zdroje patří analýza klíčových slov, která je prováděna jednou ročně. Aby nedocházelo ke ztrátě předmětové informace při zařazení klíčového slova do hesláře, dochází k automatické reindexaci bibliografických záznamů. O tu se stará nástroj Sikuli, který nahrazuje lidskou práci automatizací úkonů v grafickém rozhraní knihovnického systému Aleph.

Nejklasičtější, avšak také nejméně využívanou cestou aktualizace zůstává formulář pro návrh nového hesla.

Za účelem usnadnění práce a zvýšení atraktivity byly vytvořeny nové nástroje na prohlížení PSH, které jsou rozděleny podle skupiny jeho uživatelů.

Pro aktivní uživatele, byl vytvořen nový PSH Manager. Ten klade důraz na přehledné zobrazení struktury, obsahovou aktuálnost a implementuje funkční požadavky katalogizátorů. Je přístupný nejen pracovníkům NTK, ale také indexátorům spřízněných institucí.

Pro běžné uživatele jsou v pokročilých fázích vývoje dvě z reprezentací PSH. V rámci stránek Národní technické knihovny se jedná o tag cloud graficky zachycující aspekt využití jednotlivých hesel v knihovním fondu.

Původní Prohlížení PSH prochází řadou grafických a funkčních změn. Zásadní bylo jeho přesunutí na samostatné webové stránky, které prostorově neomezuji

jednotlivé části. Nově jsou přidávány odkazy na namapovaná hesla ve Wikipedii a současně s integrací našeptávače hesel PSH.

Heslář přestal být využíván pouze při manuální indexaci dokumentů, ale byl také integrován jako součást procesu automatické indexace v Národním úložišti šedé literatury. Tento proces je semiautomatický a uživatel nad ním má plnou kontrolu. Může tedy o použití indexace i o výsledných heslech rozhodovat.

Dalším nevšedním způsobem využití je kategorizace pracovníků Univerzitní knihovny v Pardubicích, která si PSH sama vyhledala a aktivně přispívá k jeho doplňování a opravám.

10 Seznam použité literatury

1. BROŽEK, A. Alexandrijská knihovna. *Knihovnická Revue* [online]. 2002, no. 2 [cited 2011-04-07], p. 122–124. Available from <http://knihovna.nkp.cz/Nkkr0202/0202122.html>. ISSN 1214-0678.
2. BINDING, C., TUDHOPE, D. Terminology Web Services. *Knowledge organization* [online]. 2010, vol. 37, no. 4 [cited 2011-03-15], p. 287–298. Available from http://www.ergon-verlag.de/isko_ko/downloads/ko37_2010_4_e.pdf. ISSN 0943-7444.
3. HODGE, G. *Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files*. 1st ed. Washington D.C.: The Digital Library Federation, 2000. 43 p. ISBN 1-887334-76-9.
4. BROUGHTON, V. *Essential Thesaurus Construction*. 1st ed. Facet Publishing, 2006. 296 p. ISBN 978-1856045650.
5. MACHÁČKOVÁ, Tereza. 2006. Možnosti uspořádání fondu v Národní technické ČR. *Ikaros* [online]. 2006, roč. 10, č. 1 [cited 2006-07-11]. Available from <http://www.ikaros.cz/node/2097>. ISSN 1212-5075.
6. Knowledge organization. In *Wikipedia* [online]. Last modified on 10 February 2011 at 03:17 [cited 2011-04-12]. Available from http://en.wikipedia.org/wiki/Knowledge_organization.
7. CRAVEN, T. Thesaurus Construction. [online]. 1998 [cited 2011-04-07]. Available from <http://publish.uwo.ca/~craven/677/thesaur/main00.htm>.
8. JING, Y., CROFT, W. B. An Association Thesaurus for Information Retrieval. In *In RIAO 94 Conference Proceedings*. 1994, p. 146–160.
9. JANSEN, B. J. Search log analysis: What it is, what's been done, how to do it. *Library & Information Science Research*, 2006, vol. 28, p. 407–432. ISSN 1394-6234.

10. SKOLKOVÁ, L. *Polytematický strukturovaný heslář: Diplomová práce*. Praha: Univerzita Karlova v Praze, 23.4.2007. 172, 11 p.
11. KLOUČKOVÁ, Zdenka. 1997. Polytematický strukturovaný heslář Státní technické knihovny. Čtenář. 1997, roč. 49, č. 4, s. 128-129. ISSN 0011-2321.
12. HAUZNER, Ivan. 2005. PSH - Polytematický strukturovaný heslář. Knihovnický pravodaj Vysočina [online]. 2005, roč. 6, č. 3 [cited 2007-02-23]. Available from <http://kzv.kkvysociny.cz/Default.aspx?id=190>. ISSN 1213-8231.
13. ŠEVČÍKOVÁ, B. Zápis z 8. zasedání Rady pro koordinaci Polytematického strukturovaného hesláře(PSH) ze dne 11.12. 2006. [online]. 2006 [cited 2011-03-09]. Available from www.techlib.cz/files/download/id/144/Zápis%20z%20porady%202006.pdf.
14. CAMPOS, F. M., LOPES, M. I., GALVAO, R. M. MARC formats and their use: An overview. *Program: Automated Library and Information Systems*, 1995, vol. 29, no. 4, p. 445–459. ISSN 0033-0337.
15. HOPKINSON, A. MARC developments in UNITED KINGDOM. *University Library E-Sources INFOTHECA* [online]. 2003 [cited 2011-04-09]. Available from <http://www.libraryjournal.com/article/CA250046.html>.
16. MARC 21 format for bibliographic data. *MARC Standards* [online]. [cited 2011-04-07]. Available from <http://www.loc.gov/marc/>.
17. Švástová, P. Konverze záznamů z MARCXML do MODS a MADS. diplomová práce, Masarykova Univerzita, 2006.
18. Library of Congress. 2006. MARC 21 XML Schema [online]. Washington (DC, USA) The Library of Congress, July 26, 2006 [cited 2007-04-05]. Available from <http://www.loc.gov/standards/marcxml/>.
19. Extensible Markup Language (XML). [online]. [cited 2011-04-07]. Available from <http://www.w3.org/XML/>.

20. TENNANT, R. MARC Must Die. *Library Journal* [online]. 2002 [cited 2011-04-09]. Available from <http://www.libraryjournal.com/article/CA250046.html>.
21. MYNARZ, J., ZEMÁNEK, J. Úvod k linked data. *Knihovna plus* [online]. 2010, no. 1 [cited 2011-04-07]. Available from <http://knihovna.nkp.cz/knihovnaplus101/myna.htm>. ISSN 1801-5948.
22. Resource Description Framework (RDF). [online]. 2004 [cited 2011-04-03]. Available from <http://www.w3.org/RDF/>.
23. MILES, A., MATTHEWS, B., WILSON, M., BRICKLEY, D. SKOS Core: Simple Knowledge Organisation for the Web. [online]. 2005 [cited 16-04-11]. Available from <http://epubs2.cclrc.ac.uk/bitstream/675/dc2005skospaperssubmission1.pdf>.
24. World Wide Web Consortium (W3C). [online]. [cited 2011-03-10]. Available from <http://www.w3.org/>.
25. VAN ASSEM, M., et al. A method to convert thesauri to SKOS. *The Semantic Web: Research and Applications*. 2006, vol. 4011 [cited 27-03-11], p. 95–109.
26. TORDAI, A., et al. Aligning Large SKOS-Like Vocabularies: Two Case Studies. In *Proceedings of ESWC: Extended Semantic Web Conference*. 2010, p. 198–212.
27. A. Miles, N. Rogers, and D. Beckett. Migrating Thesauri to the Semantic Web - Guidelines and case studies for generating RDF encodings of existing thesauri. Deliverable 8.8, SWAD-Europe [online]. 2004 [cited 2011-04-16]. Available from <http://www.w3.org/2001/sw/Europe/reports/thes/8.8/>.
28. BERNERS-LEE, T. Linked Data. *International Journal on Semantic Web and Information Systems* [online]. 2006, vol. 4, no. 2 [cited 2011-04-16].

Available from <http://www.w3.org/DesignIssues/LinkedData.html>. ISSN 1552-6283.

29. Bizer, C.; Cyganiak, R.; Heath, T. How to Publish Linked Data on the Web [online], 2007 [cited 2011-04-16]. Available from <http://www4.wiwiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/>.
30. BIZER, C., HEATH, T., BERNERS-LEE, T. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*[online]. 2009. Available from <http://eprints.ecs.soton.ac.uk/21285/1/bizer-heath-berners-lee-ijswis-linked-data.pdf>.
31. BERNERS-LEE, Tim; HENDLER, James; LASSILA, Ora. 2001. The Semantic Web : A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American*. May 17, 2001. [cited 2003-11-06]. Available from <http://www.scientificamerican.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21&catID=2>.
32. PASTOR-SANCHEZ, J. A., MENDES, F. J. M., RODRIGUEZ-MUNOZ, J. V. Advantages of thesaurus representation using the Simple Knowledge Organization System (SKOS) compared with proposed alternatives. *Information Research: an international electronic journal*[online]. 2009, vol. 14, no. 4 [cited 2011-04-02], p. 246–250. Available from <http://informationr.net/ir/14-4/paper422.html>. ISSN 1545-598X.
33. SCHANDL, T., BLUMAUER, A. PoolParty: SKOS Thesaurus Management Utilizing Linked Data Heraklion, GREECE, MAY 30-JUN 03, 2010. In AROYO, L., et al. (ed.). *7th Extended Semantic Web Conference (ESWC2010) Heraklion, GREECE, MAY 30-JUN 03, 2010: Semantic web: Research and applications, PT 2, Proceedings*. 2010, p. 421–425.
34. Věcné autority. *Národní autority ČR* [online]. [cited 2011-03-10]. Available from <http://authority.nkp.cz/vecne-autority>.

35. MYNARZ, J. Jak lze prakticky využít Polytematický strukturovaný heslář pro věcný popis elektronických zdrojů. *Ikaros* [online]. 2009, vol. 13, no. 12 [cited 2011-04-16]. Available from <http://www.ikaros.cz/node/5872>. ISSN 1212-5075.
36. The Dublin Core® Metadata Initiative. [online]. [cited 2011-03-16]. Available from <http://dublincore.org/>.
37. Common Tag. [online]. [cited 2011-03-16]. Available from <http://commontag.org/Home>.
38. KOŽUCHOVÁ, K., ŠKUTA, C. 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. května 2010* [online]. Praha: Albertina icome Praha, 2010, [cited 2011-04-16]. Available from <http://www.inforum.cz/sbornik/2010/67>. ISSN 1801-2213.
39. *Python Programming Language* [online]. 1990 [cited 2011-04-27]. Official Website. Available from <http://python.org/>.
40. PILGRIM, M. *Dive Into Python*. 1st ed. CreateSpace, 2009. 330 p. ISBN 978-1441413024.
41. *Django* [online]. 2005 [cited 2011-04-27]. The Web framework for perfectionists with deadlines. Available from <http://www.djangoproject.com/>.
42. KAPLAN-MOSS, J., HOLOVATY, A. *The Definitive Guide to Django: Web Development Done Right*. 2nd ed. Apress, 2009. 536 p. ISBN 978-1430219361.
43. HASSAN, S. S., ISAAK, R. K. An integrated approach of MAS-CommonKADS, Model-View-Controller and web application optimization strategies for web-based expert system development. *Expert Systems with Applications*, 2011, vol. 38, no. 1, p. 417–428. ISSN 0957-4174.

44. JavaScript. *MDN Docs* [online]. 2011 [cited 2011-04-05]. Available from <https://developer.mozilla.org/en/JavaScript>.
45. RESIG, J., et al. *jQuery* [online]. 2010 [cited 2011-04-27]. Available from <http://jquery.com/>.
46. MySQL. [online]. 2011 [cited 2011-04-05]. Available from <http://www.mysql.com/>.
47. Ragget, D.; et al. HTML 4.01 Specification. 1999 [cited 2011-04-16]. Available from <http://www.w3.org/TR/html401/>.
48. BOS, B., et al. (eds.). *Cascading Style Sheets Level 2 Revision 1 (CSS 2.1) Specification* [online]. 2011-04-12 [cited 2011-04-16]. Available from <http://www.w3.org/TR/CSS21/>.
49. ZHONG, L., LUANG, S. S., LIU, P., et al. A design for an online RSS reader based on AJAX. In FENG, W. Y., GAO, F. (ed.). *8th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing/3rd ACIS International Workshop on Self-Assembling Wireless Networks Qungdao, PEOPLES R CHINA, JUL 30-AUG 01, 2007: SNPD 2007: Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, Vol 1, Proceedings*. 2007, p. 794–798. ISBN 978-0-7695-2909-7.
50. JUVAN, S., BARTOL, B., BOH, T. Data structuring and classification in newly-emerging scientific fields. *Online Information Review*, 2005, vol. 29, no. 5, p. 483–498. ISSN 1468-4527.
51. HOCKE, R., CHANG, T. H. The Complete Guide to Sikuli Script. *Project Sikuli* [online]. 2009 [cited 2011-03-04]. Available from <http://sikuli.org/trac/wiki/reference-0.10#TheCompleteGuidetoSikuliScript>.
52. MASCHOTTA, R., BOYMANN, S., HOPPE, U. Comparison of feature-list cross-correlation algorithms with common cross-correlation algorithms. *EURASIP*

- Journal on Advances in Signal Processing*[online]. 2007 [cited 2011-04-06], p. 89150. Available from http://www.tu-ilmenu.de/fakia/fileadmin/template/startIA/bmti_neu/publications/2007/Maschotta_Comparison_of_Feature-List_Cross-Correlation_Algorithms_with_Common_Cross-Correlation_Algorithms_komplett.pdf. ISSN 1687-6172 .
53. MACH, L. SIFT: Automatické nalezení korespondencí mezi dvojicí obrázků,. [online]. 2007 [cited 2011-03-09]. Available from <http://cgg.mff.cuni.cz/~pepca/ref/SIFT.pdf> .
54. LOWE, D. G. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* [online]. 2004, vol. 60, no. 2 [cited 2011-04-04], p. 91–110. Available from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.157.3843&rep=rep1&type=pdf> . ISSN 0920-5691.
55. CHENG, L. A., GONG, J. Y., YANG, X. X. Robust affine invariant feature extraction for image matching. *IEEE Xplore - Geoscience and Remote Sensing Letters* [online]. 2008, vol. 5, no. 2 [cited 2011-04-12], p. 246–250. Available from <http://rcv.kaist.ac.kr/~shkim/document/Zhe%20Lin.pdf> . ISSN 0920-5691.
56. BATTY, D. WWW -- Wealth, Weariness or Waste - Controlled vocabulary and thesauri in support of online information access. *D-Lib Magazine* [online]. 1998. Available from <http://www.dlib.org/dlib/november98/11batty.html> . ISSN 1082-9873.
57. HOCHSTOTTER, N., KOCH, M. Standard parameters for searching behaviour in search engines and their empirical evaluation. *Journal of Information Science*, 2007, vol. 35, no. 1, p. 45–65. ISSN 0165-5515.
58. HODKINSON, C., KIEL, G., McCOLL-KENNEDY, J. R. Consumer web search behaviour: diagrammatic illustration of wayfinding on the web. *International Journal of Human-Computer Studies*, 2000, vol. 52, no. 5, p. 805–830. ISSN 1071-5819.

59. JONES, S., et al. A Transaction Log Analysis of a Digital Library. *International Journal on Digital Libraries* [online]. 2000, vol. 3 [cited 2011-04-17], p. 152–169.
60. JANSEN, B. J. Search log analysis: What it is, what's been done, how to do it. *Library & Information Science Research*, 2006, vol. 28, p. 407–432. ISSN 1394-6234.
61. CHAU, M., FANG, X., SHENG, O. R. L. Analysis of the query logs of a web site search engine. *Journal of the American Society for Information Science and Technology* [online]. 2005, vol. 56 [cited 2011-03-06], p. 1363–1367. Available from <http://www.fbe.hku.hk/~mchau/papers/QueryLogs.pdf>. ISSN 1532-2882.
62. RAVID, G., BAR-ILAN, J., BARUCHSON-ARBIB, S. Popularity and findability through log analysis of search terms and queries: the case of a multilingual public service website. *Journal of Information Science* [online]. 2007, vol. 33, no. 5 [cited 2011-04-04], p. 567–583. Available from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.99.5899&rep=rep1&type=pdf>. ISSN 0165-5515.
63. SINCLAIR, J., CARDEW-HALL, M. The folksonomy tag cloud: when is it useful?. *Journal of Information Science* [online]. 2008, vol. 34, no. 1 [cited 2011-03-02], p. 15–29. Available from <http://www.jrsinclair.com/academic/JIS-0449%20-v3%20-%20The%20Folksonomy%20Tag%20Cloud%20-%20When%20is%20it%20useful.pdf>. ISSN 0165-5515.
64. *The Apache Xalan Project* [online]. [cited 2011-04-17]. Available from <http://xalan.apache.org/>
65. GARRET, J. J. Ajax: A New Approach to Web Applications. [online]. 2005 [cited 2011-04-09]. Available from <http://www.adaptivepath.com/ideas/e000385>.

66. HTTP - Hypertext Transfer Protocol. [online]. 2011 [cited 2011-04-14]. Available from <http://www.w3.org/Protocols/> .
67. KIPP, M., CAMPBELL, D. G. Searching with Tags: Do Tags Help Users Find Things?. *Knowledge organization* [online]. 2010, vol. 37, no. 4 [cited 2011-03-20], p. 239–255. Available from http://www.ergon-verlag.de/isko_ko/downloads/ko37_2010_4_a.pdf. ISSN 0943-7444.
68. MYNARZ, J., ŠKUTA, C. Integration of an Automatic Indexing System within the Document Flow of a Grey Literature Repository. In *12th International Conference on Grey Literature, DEC 06-07, 2010 Natl Techn Lib, Prague, CZECH REPUBLIC: TRANSPARENCY IN GREY LITERATURE: GREY TECH APPROACHES TO HIGH TECH ISSUES*. 2011, p. 65–71. ISBN 978-90-77484-16-6.
69. BATES, M. J. Indexing and access for digital libraries and the Internet: Human, database, and domain factors. *Journal of the American Society for Information Science and Technology*, 1998, vol. 49, no. 13, p. 1185–1205. ISSN 1532-2890.
70. BLACKLER, F. Knowledge and the theory of organizations as activity systems and the reframing of management. *Journal of Management Studies*, 1993, vol. 30, no. 6, p. 863–884. ISSN 0022-2380.
71. PEPE, A. [et al.]. CERN Document Server Software : the integrated digital library. In DOBREVA, Milena; ENGELEN, Jan (eds.). *9th ICC International Conference on Electronic Publishing : from author to reader : challenges for the digital content chain*. Leuven : Peeters, 2005. ISBN 90-429-1645-1.
72. Invenio. [online]. 2011 [cited 2011-04-07]. Available from <http://invenio-software.org/> .
73. MEDELYAN, Olena. *Human-competitive automatic topic indexing*. Waikato, 2009. 214 s. Disertační práce (PhD.). University of Waikato, Department of computer science, 2009.

74. MEDELYAN, Olena. *Useful web resources related to automatic topic indexing* [online]. July 13, 2009 [cited 2009-11-11]. Available from <http://maui-indexer.blogspot.com/2009/07/useful-web-resources-related-to.html>
75. KUČERA, Karel. The Czech National Corpus : principles, design, and results. *Literary and Linguistic Computing*. 2002, vol. 17, no. 2, p. 245 –257. ISSN 0268-1145.
76. DOLAMIC, Ljiljana; SAVOY, Jacques. Indexing and stemming approaches for the Czech language. *Information Processing & Management*. November 2009, vol. 45, iss. 6, p. 714 – 720. ISSN 0306-4573.
77. PEJŠOVÁ, Petra (ed.). *Grey literature repositories*. Zlín: VeRBuM, 2010, 156 p. ISBN 978-80-904273-6-5.
78. SHAH, P., PEREZ-IRATXETA, C., BORK, P., et al. Information extraction from full text scientific articles: Where are the keywords?. *BMC Bioinformatics* [online]. 2003, vol. 4 [cited 2011-04-26], p. 4. Available from <http://www.biomedcentral.com/1471-2105/4/20> . ISSN 1471-2105.
79. BRACEWELL, D. B., YAN, R., REN, F. Single document keyword extraction for Internet news articles. *International Journal of Innovative Computing, Information and Control*, 2008, vol. 4, no. 4, p. 905–913. ISSN 1349-4198.

11 Přílohy

Vzhledem k povaze vytvořených materiálů – převážně programového kódu, i poměrně rozsáhlých vstupních datových struktur, jsou všechny přílohy obsaženy na přiloženém DVD.

Najdete zde projekty jednotlivých aplikací, které jsou rozděleny do samostatných adresářů, upravený indexační program Maui indexer, kratší jednorázové programy používané při čištění zpracovávaných dat a v neposlední řadě současnou verzi PSH ve formátu SKOS.

Připojeny nejsou pracovní databáze či jednotlivé databázové tabulky. Jejich struktura a použití vychází z programového kódu. Při zájmu je však možné tato data poskytnout.