POLYTEMATICKÝ STRUKTUROVANÝ HESLÁŘ A JEHO POTENCIÁL V OBLASTI TŘÍDĚNÍ A ZPŘÍSTUPŇOVÁNÍ WEBOVÝCH DOKUMENTŮ

Kristýna Kožuchová (kristyna.kozuchova@techlib.cz)
Národní technická knihovna

Ctibor Škuta (ctibor.skuta@techlib.cz)
Národní technická knihovna

INFORUM 2010: 16. ročník konference o profesionálních informačních zdrojích Praha, 25.–27. 5. 2010

ABSTRAKT

PSH je jedním ze zástupců systémů organizace znalostí a je proto určen k předmětové indexaci dokumentů tak, aby mohly být později efektivněji vyhledány. Zatímco v minulosti byla předmětová indexace dokumentů doménou profesionálů – knihovníků, nyní se tento druh intelektuální činnosti přesouvá také do rukou běžných uživatelů. V roce 2009 došlo v oblasti PSH k několika významným změnám. Heslář je nyní zveřejněn v sémantickém formátu Simple Knowledge Organisation System (SKOS) v souladu s principy linked data pod licencí Creative Commons. SKOS slouží k reprezentaci systémů organizace znalostí a je postaven na W3C standardech RDF (Resource Description Framework) a RDFS (RDF - Schema) za účelem zpřístupnění strukturovaných slovníků pro sémantický web. Vzhledem k používání globálně unikátních URI (Uniform Resource Identifier) jako identifikátorů je možné na hesla PSH jednoduše odkazovat a vytvářet vazby s dalšími informačními zdroji. V případě PSH to znamená, že u některých hesel lze nalézt odkazy na předmětová hesla Kongresové knihovny (Library of Congress Subject Headings) a DBPedii. Jednou z novinek zpřístupněných na stránkách Národní technické knihovny v sekci PSH je funkce pro vytváření úryvků metadat s hesly PSH ve formátech Dublin Core a Common Tag. Prostřednictvím zápisu RDF v atributech (RDFa) je možné jejich zakomponování přímo do těla dokumentů ve formátu (X)HTML bez vlivu na výsledné zobrazení. V současnosti PSH neslouží pouze jako prostředek pro věcnou selekci bibliografických záznamů, ale jeho potenciál je rozvíjen také v oblasti třídění a zpřístupňování webových dokumentů.

ABSTRACT

Polythematic Structured Subject Heading System (PSH) is an example of knowledge organisation system (KOS) which is intended for indexing and searching documents by subject. Traditionally, the subject indexing has been the responsibility of professionals - librarians. Nowadays, this type of intellectual process is also made available to general public. In 2009 PSH was published in Simple Knowledge Organisation System (SKOS) format under Creative Commons (CC) licencing terms. SKOS is designed to represent traditional knowledge organisation systems. It provides means for expressing standard relationships that can be found in most thesauri or subject heading systems. SKOS is defined using RDF Schema and expressed in RDF data model. There is

a possibility to display tags from PSH in Dublin Core and CommonTag formats in PSH section on the National technical library´s websites. Metadata tags can be embedded in (X)HTML document to achieve its semantic description. If available, there are also links reffering to similar concepts found in Library of Congress Subject Headings or on DBPedia. Currently PSH is not only a tool for bibliographic records selecting but its potential is also developed in the area of indexing and searching web documents.

KLÍČOVÁ SLOVA

Polytematický strukturovaný heslář, PSH, systémy organizace znalostí, SKOS, metadata, folksonomie

ÚVOD

Polytematický strukturovaný heslář je česko-anglický řízený slovník lexikálních jednotek, který ve své poslední verzi PSH 2.1 obsahuje přes 13 500 hesel. Je rozdělen do 44 tematických skupin, v nichž jsou hesla dále hierarchicky členěna. Ve srovnání s klíčovými slovy či nekontrolovanými termíny, které jsou vytvářeny volně, má PSH pevnou strukturu i obsah. Pravidla pro tvorbu, správu a údržbu hesláře jsou podobná pravidlům pro konstrukci tezaurů. Heslář si mohou uživatelé zdarma stáhnout a používat v souladu s podmínkami stanovenými licencí Creative Commons.

TAXONOMIE, FOLKSONOMIE, TAGOVÁNÍ

PSH je jedním ze zástupců systémů organizace znalostí a je proto určen k předmětové indexaci dokumentů tak, aby mohly být později efektivněji vyhledány. Zatímco v minulosti byla předmětová indexace dokumentů doménou profesionálů – knihovníků, nyní se tento druh intelektuální činnosti přesouvá také do rukou běžných uživatelů. Vedle klasických forem vyhledávání a hierarchického kategorického třídění informací se v rámci organizace webového obsahu můžeme setkat s poměrně novým typem třídění, které si vytvářejí sami uživatelé - folksonomie.

Folksonomie je na rozdíl od taxonomie představující předem vytvořenou kategorizaci pojmů tvořena volně bez daných pravidel, hierarchie a řádu, na níž lze nahlížet jako na metodu volného třídění informací s použitím tagů. Folksonomie byly rovněž popsány jako uživatelem vytvářená metadata. Tagy si můžeme představit jako jednotlivá klíčová slova, která volně vytváří sami uživatelé. Při tvorbě tagů nejsou svázáni žádnými pravidly, tagy tvoří intuitivně, v přirozeném jazyce. Co se týče významu a důležitosti jsou jednotlivé tagy na stejné úrovni, čímž vzniká tzv. heterarchie, tedy soustava rovnocenných prvků.

Vzhledem k tomu, že je třídění pomocí tagů v poslední době velice populární, existuje řada webových služeb, které nabízí nejrůznější funkce personalizovaného třídění dokumentů. Mezi nejznámější patří Delicious a Digg pro širokou veřejnost nebo Connotea a CiteULike určené výhradně pro odborné články. Mezi našimi zástupci bychom mohli uvést například službu Bookmarky.cz. Folksonomie jsou využívány také v komerčním prostředí. Příkladem je internetový obchod Amazon, který je proslulý pro nabídkové funkce svého katalogu. Jeho zákazníci mohou označovat zboží podle svého zájmu uživatelskými tagy. Jejich přidávání je na

Amazonu podmíněno přihlášením uživatele do systému, které slouží jako ochrana proti spamu. Výsledné tagy pak mohou být zpřístupněny lidem s podobnými zájmy.

V oblasti uživatelského tagování existuje také přístup, který spočívá v poskytnutí souboru termínů povolených k přiřazení, ideálně z řízeného slovníku. Uživatel si v tomto případě vybere vhodný termín z přiloženého seznamu, čímž zároveň zařazuje dokument do dané hierarchie, určuje dokumenty podobné, spadající do stejného oboru, spolu s dokumenty obecnějšími nebo specifičtějšími. Problémy spojené s přirozeným přístupem k uživatelskému tagování jsou tedy vyřešeny, vznikají však jiné. Volnost tvorby termínů se ztrácí, společně se svobodou a pohodlím uživatele v přiřazování hesel. Důležitá je také velikost přiloženého hesláře. Malý počet hesel indexaci usnadňuje, ale neumožňuje přesnější zařazení dokumentu. Příliš velké množství naopak prodlužuje dobu indexace a odrazuje od hledání vhodného hesla.

Na jedné straně tedy máme možnost pohodlného tagování v přirozeném jazyce bez nutnosti výběru vhodného termínu z omezeného seznamu, avšak bez dané struktury a s velmi širokou a nejednoznačnou základnou pojmů. Řízený slovník naopak nabízí vytvořenou strukturu a provázanost jednotlivých pojmů, na druhou stranu není výběr klíčových slov pohodlný a někdy je velmi těžké najít ten "správný" termín.

Existují projekty, které se snaží spojit výhody obou přístupů, tedy uživatelsky tvořených hesel, ale s definovanou strukturou. Jedním z nich je například SOBOLEO (Social Bookmarking and Lighweight Engineering of Ontologies), které sice používá řízený slovník ve formátu SKOS (Simple Knowledge Organisation System), ale jeho tvorbu nechává na samotných uživatelích. Ti tak mohou vkládat hesla do struktury slovníku, včetně určení synonym v daném i cizích jazycích nebo provázání s příbuznými termíny.

Podobný princip by bylo možné uplatnit také v knihovnách. Automatizované knihovnické systémy jsou obvykle vybaveny sofistikovaným vyhledávacím rozhraním, které pracuje s nejrůznějšími systémy organizace znalostí. Prakticky vždy jsou však dokumenty indexovány odborníky. Uživatelské rozhraní poskytuje pouze omezené (či žádné) možnosti ukládání a následného třídění záznamů. Polytematický strukturovaný heslář by se tak mohl využít pro sémantické rozšíření vyhledávání dokumentů v katalogu NTK o funkci tagování a třídění záznamů. Uživatelé by měli možnost tagovat jednotlivé záznamy, zařazovat je do složek nebo definovat obory svého zájmu. Tagy by mohly být vybírány z hesláře, ale každý uživatel by si mohl vytvářet i své vlastní. S pomocí PSH by také mohl definovat oblasti svého zájmu, podle kterých by následně dostával upozornění na nové tituly. Tematické zaměření čtenářů by navíc umožňovalo navrhování knih mezi uživateli se stejnými zájmy. Vyhledávání by se poté stále primárně opíralo o bibliografické záznamy s možností rozšíření o uživatelské značky, ale bylo by možné zobrazit i tituly označené daným heslem/tagem.

GENEROVÁNÍ ÚRYVKŮ METADAT S HESLY PSH

Metadata, jejichž využití pro zachycení a vyjádření sémantiky jsou nezbytným předpokladem pro fungování sémantického webu, jsou další oblastí, kde dochází k přesunu předmětové indexace dokumentů z rukou knihovníků tentokrát do rukou autora dokumentu. Tvorbu metadat založených na daném metadatovém formátu usnadňují různé generátory úryvků metadat (například generátor na stránkách WebArchivu), které pomáhají uživateli zejména s celkovou syntaxí metadatového záznamu. K nejznámějším formátům patří pravděpodobně Dublin Core. Novou funkcí zpřístupněnou na stránkách hesláře je generování úryvků metadat s hesly PSH. Tímto způsobem je možné získat tagy ve formátech Dublin Core nebo CommonTag. Prostřednictvím zápisu RDF v atributech (RDFa) je možné jejich zakomponování přímo do těla dokumentů ve formátu (X)HTML bez vlivu na výsledné zobrazení. Metadata ve formátech Dublin Core a Common Tag lze jednoduše vložit do webových stránek k dosažení jejich sémantického popisu.

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. V současnosti se jedná patrně o nejrozšířenější způsob zápisu metadat.

CommonTag je poměrně mladý sémantický formát (srpen, 2009), určený pro popis (X)HTML dokumentů. Využívá standardu RDFa (RDF v atributech), tedy standartního způsobu pro zápis metadat, který je podporován oběma největšími hráči na poli vyhledávání, společnostmi Google a Yahoo!. Yahoo! byl také jednou z několika velkých firem, které se podíleli na jeho vývoji.

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/PSH6185"
rel="dc:subject">biochemie</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/PSH6185"
property="ctag:label" content="biochemie" />
</body>
```

Obr. 1 Úryvky metadat s hesly PSH

ANALÝZA UŽIVATELSKÝCH DOTAZŮ

V minulém roce byly zahájeny přípravy aktualizace současné verze PSH 2.1. V této souvislosti byl vytvořen jednoduchý webový formulář pro zasílání návrhů nového hesla PSH. Využívat jej mohou jak katalogizátoři oddělení věcného zpracování v NTK, tak externí uživatelé, např. zájemci o problematiku selekčních jazyků.

Formulář pro zasílání návrhů nového hesla PSH je využíván také zákazníky NTK. Jedná se však spíše o ojedinělé případy. Představu o tom, podle jakých kritérií uživatelé vyhledávají dokumenty v katalogu NTK a do jaké míry se liší od hesel PSH, lze získat analýzou logu uživatelských dotazů zadávaných uživateli při vyhledávání v katalogu NTK, kterou se referát PSH soustavně zabývá od počátku letošního roku.

„Log soubory, zkráceně logy, jsou textové soubory obsahující záznamy o činnosti nějaké konkrétní aplikace. V případě webových serverů jsou do logů ukládány veškeré požadavky, které byly na server vzneseny. Zpětnou analýzou těchto dat pak můžeme zjišťovat cenné informace o fungování sledovaného webu“ (SMRT, 2007). Analýza webových logů je nejčastěji spojována se statistikou počtu návštěvníků, průměrné doby jejich setrvání, statistiky přístupů apod. K výhodám analýzy logů patří možnost jejího jednorázového provedení za delší období a fakt, že větší objem dat poskytne statisticky významnější výsledky.

Cílem rozboru logu uživatelských dotazů v referátu PSH v NTK je optimalizace aktualizace PSH, která by měla usnadnit uživatelům vyhledávání a přiblížit se jejich potřebám. Rozbor dotazů pokládaných uživateli při práci v online katalogu může do určité míry vyřešit disproporci mezi termíny používanými při indexaci a termíny používanými samotnými uživateli popisujícími jejich informační potřeby.

Pro analýzu logu v NTK se používá skript napsaný v programovacím jazyce Python. Po načtení do paměti je text(log) sekvenčně prohledán pro shodu s určeným regulárním výrazem. Regulární výrazy představují uživatelsky navržené šablony textu, proto je jejich použití vhodné pro práci s textem s definovanou - opakující se strukturou. Vybírány jsou požadavky (requesty) vyhledávající ve všech polích, v polích pro předmětová hesla nebo hesla PSH. Automaticky jsou vyloučeny termíny obsahující číslice (ISBN, ISSN aj.). Vliv násobného odeslání požadavku jednotlivých uživatelů je filtrován na základě časových a identifikačních údajů, tedy data, času a IP adresy. K tomuto jevu dochází při návratu na stránku s výsledkem hledání v katalogu NTK z prohlíženého záznamu. Vyextrahované termíny jsou následně porovnány se současnými hesly PSH za vytvoření informační statistiky jejich využitelnosti. Zbývající termíny, které jsou potencionálními kandidáty na nové deskriptory/nedeskriptory hesláře, přidáváme do výstupního souboru společně se sumární hodnotou jejich použití.

Vzhledem k faktu, že uživatelé v majoritním počtu případů používají vyhledávání ve všech polích, výsledný soubor stále obsahuje velké množství na první pohled jasně nevyhovujících výrazů (jména autorů, nakladatelství). O konečný výběr vyhovujících termínů se již při ručním zpracování starají pracovníci referátu PSH společně s oddělením správy autoritních souborů. Posuzování probíhá ve směru klesající frekvence výskytu jednotlivých slov s nastavenou spodní prahovou hodnotou. Sledování uživatelského chování je velmi užitečné, jelikož názorně ukazuje, jakým způsobem se uživatelé v katalogu NTK pohybují, jaký postup uplatňují při hledání určitého dokumentu a jak se jim to daří.

SKOS (SIMPLE KNOWLEDGE ORAGANIZATION SYSTEM)

Podoba webu dramaticky ovlivnila aplikaci tezaurů. Samotný proces integrace různých tezaurů s informačními vyhledávacími systémy začal v 90. letech minulého století. Ačkoli se některé z nich adaptovaly do digitálního prostředí, jejich efektivní využití na webu je oproti snaze stále malé mezi jinými je to také z těchto důvodů:

- Limitovaný vývoj a vymezení konceptuálně orientovaných tezaurů. Aplikace lexikálních tezaurů v dynamickém prostředí jako je web neposkytuje adekvátní výsledky ve srovnání se snahou věnovanou indexačnímu procesu.
- Počáteční absence vhodných standardů a modelů k reprezentaci různých stupňů abstrakce na webu (např. XML (the Extensible Markup Language), RDF (the Resource Description Framework) nebo SKOS (the Simple Knowledge Organization System)).
- Představa tezaurů jako zastaralého nástroje nebo nástroje s limitovaným použitím.

Hlavní formát, v němž je PSH zachycen, je MARC pro autoritní záznamy, který se používá v automatizovaných knihovnických systémech, ale pro webovou distribuci se nehodí. Součást procesu "webifikace" tezaurů zahrnuje převod existujících nástrojů do sémantických webových standardů jako je SKOS nebo OWL. Standardizace a globální užití formátu MARC umožňuje tyto slovníky převádět do přívětivých RDF dat.

SKOS je jednoduchý formát určený pro reprezentaci, sdílení a odkazování znalostních systémů. Do této skupiny patří tezaury, klasifikační schémata, předmětová hesla, kontrolované slovníky představující soubory pojmů potenciálně zahrnující údaje o vztazích mezi těmito pojmy. Jejich společným rysem je jejich určení k užití v informačních vyhledávacích aplikacích. SKOS je založen na standardech konsorcia W3C RDF (Resource Description Framework) a RDFS (RDF Schema) a vzhledem k tomu je možné jej používat v kombinaci s dalšími RDF formáty.

Inspirací pro převod a způsob distribuce PSH ve formátu SKOS byly již existující aplikace jazyka SKOS v jiných knihovnách. Jako jedna z prvních knihoven použila SKOS pro své věcné autority Švédská národní knihovna. Jiným příkladem je Tezaurus pro ekonomii Německé národní ekonomické knihovny. V květnu 2009 byla v tomto formátu oficiálně zveřejněna také předmětová hesla Kongresové knihovny (Library of Congress Subject Headings). Data PSH ve formátu SKOS jsou k dispozici jako dávkový export, který je volně ke stažení pod podmínkami licenčního ujednání Creative Commons CC-BY-NC-SA (Uveďte autora – Neužívejte dílo komerčně – Zachovejte licenci 3.0 Česko). Pro strojově čitelné vyjádření licence Creative Commons byl použit jazyk Creative Commons Rights Expression Language (ccREL), který je rovněž založen na RDF, a proto je možné jej začlenit do mnoha dalších formátů (XHTML, XML, SKOS aj.).

Heslář je v souladu se čtyřmi principy linked data, což v případě PSH zahrnuje:

1. Použití URI (Uniform Resource Identifier) jako identifikátorů hesel.
2. Použití HTTP (Hypertext Transfer Protocol) URI, takže je možné si hesla prohlédnout pomocí webového prohlížeče.
3. Poskytnutí užitečných informací u hesla.
4. Zahrnutí odkazů na další zdroje.

PSH/MARC používá k linkování stanovená záhlaví, zatímco SKOS koncepty (elementy skos:concept) jsou vzájemně propojeny použitím jednotných identifikátorů URI. Každý autoritní záznam dodávaný NTK zahrnuje kontrolní číslo v MARC v poli 001. Tato kontrolní čísla jsou vytvářena tak, aby byla persistentní a jedinečná. Každá kontrolní číslice se skládá z alfanumerického kódu, který je tvořen tříznakovou zkratkou „PSH“, za níž následuje unikátní číselná kombinace, např. PSH5450. Toto číslo se stalo kandidátem pro identifikaci SKOS konceptů.

příklad:

<skos:Concept rdf:about="http://psh.ntkcz.cz/skos/PSH5450">

Formát MARC pro autority rozlišuje preferovaná (1XX) a nepreferovaná (4XX) znění hesel. Podobně také SKOS slovník poskytuje dvě charakteristiky - skos:prefLabel a skos:altLabel, což umožňuje jejich přímé namapování. SKOS byl vytvořen pro multi-jazykové prostředí. Uživatelé SKOS jsou podporováni, aby užívali atributy pro určení jazyka. PSH je dvojjazyčný, proto má každé heslo českou i anglickou podobu.

příklad:

<skos:prefLabel xml:lang="cs">chemie</skos:prefLabel>
<skos:prefLabel xml:lang="en">chemistry</skos:prefLabel>

PSH/MARC používá pole 5XX pro vazby mezi jednotlivými preferovanými zněními. Sémantické vztahy vyjádřené ve formátu MARC jsou snadno převoditelné do formátu SKOS, který je definuje pomocí skos:related, skos:broader, skos:narrower.

příklad:

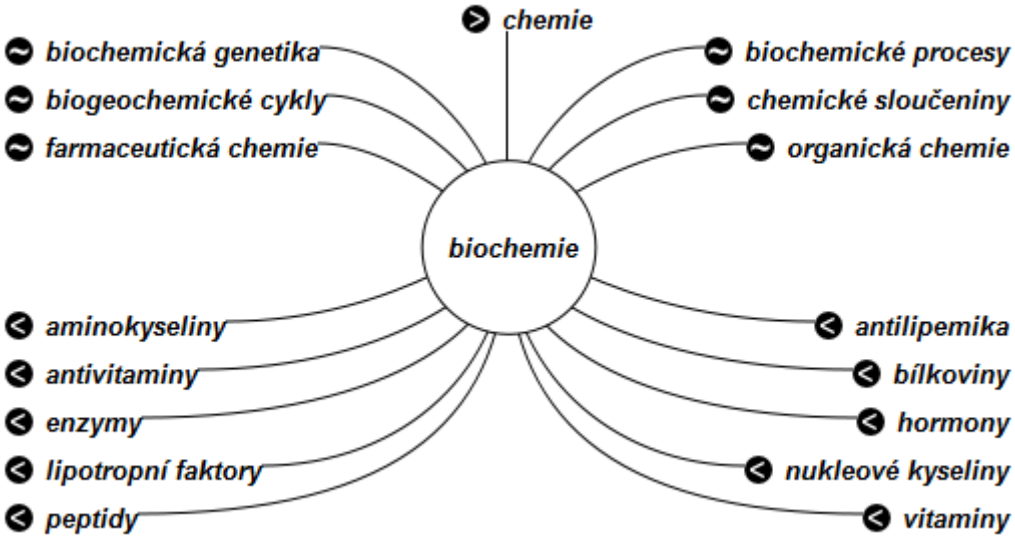
<skos:related rdf:resource="http://psh.ntkcz.cz/skos/PSH423"/>
<skos:broader rdf:resource="http://psh.ntkcz.cz/skos/PSH3423"/>
<skos:narrower rdf:resource="http://psh.ntkcz.cz/skos/PSH3474"/>

```
−<rdf:RDF>
  −<skos:Concept rdf:about="http://psh.ntkcz.cz/skos/PSH807">
      <skos:prefLabel xml:lang="cs">geny</skos:prefLabel>
      <skos:prefLabel xml:lang="en">genes</skos:prefLabel>
      <skos:broader rdf:resource="http://psh.ntkcz.cz/skos/PSH803"/>
      <skos:narrower rdf:resource="http://psh.ntkcz.cz/skos/PSH812"/>
      <skos:narrower rdf:resource="http://psh.ntkcz.cz/skos/PSH813"/>
      <skos:narrower rdf:resource="http://psh.ntkcz.cz/skos/PSH810"/>
      <skos:narrower rdf:resource="http://psh.ntkcz.cz/skos/PSH811"/>
      <skos:narrower rdf:resource="http://psh.ntkcz.cz/skos/PSH808"/>
      <skos:narrower rdf:resource="http://psh.ntkcz.cz/skos/PSH809"/>
      <skos:exactMatch rdf:resource="http://id.loc.gov/authorities/sh91000344#concept"/>
      <skos:inScheme rdf:resource="http://psh.ntkcz.cz/skos"/>
    </skos:Concept>
</rdf:RDF>
```

Obr. 2 Heslo PSH v formátu SKOS

Převedení a distribuce PSH ve formátu SKOS je přínosné z několika hledisek. SKOS je moderní formát a je tedy podporovaný řadou softwarových nástrojů z funkčně odlišných oblastí. Jedním ze základních nástrojů jsou validátory, které zajišťují automatickou revizi konsistence hesláře. Kontrolují vztahy vyplívající z logiky daného formátu, tedy správnost strukturního provázání hesel například nemá-li jedno heslo více nadřazených hesel apod. Kontrola potenciálních chyb je tak velmi snadná, rychlá a explicitní.

Vizualizační nástroje mohou převádět SKOS do jeho grafické reprezentace, která je nejen uživatelsky atraktivní, ale velmi často napomáhá v orientaci při logickém procházení struktury hesláře. Ukázkou reprezentace struktury formátu SKOS je zobrazení jednotlivých hesel PSH pod odkazem PROHLÍŽENÍ HESLÁŘE na stránkách Národní technické knihovny.



Obr. 3 Grafické zobrazení hesla PSH

Další možností je využití PSH/SKOS při automatické indexaci dokumentů s pomocí řízeného slovníku. SKOS je akceptován jako vstupní formát při indexačním procesu a výstupem v podobě dokumentu s přiřazenými předmětovými hesly. Velkou výhodou formátu SKOS je v tomto případě jeho multijazyková podpora. Při použití dobře zvoleného indexačního nástroje je možné indexovat nejen nejběžnější dokumenty anglické, ale například i texty české. A to nejen na základě statistiky slov a slovních spojení, ale i učícími algoritmy s pokročilou analýzou přirozeného jazyka.

SKOS je aplikací RDF, který dobře definuje logické vlastnosti. Díky tomu jsou kontrolované strukturované slovníky používající SKOS strojově srozumitelné např. počítačová aplikace umí číst, dávat význam a poskytovat jim různé funkce. RDF podporuje distribuované publikování dat. To znamená, že strukturované kontrolované slovníky zveřejněné ve formátu SKOS mohou být odkazovány na další datové zdroje (např. další slovníky apod.). Tuto vlastnost demonstruje namapování hesel PSH na Předmětová hesla Kongresové knihovny (LCSH) a DBPedii.

DBpedia je projekt, jehož cílem je extrakce strukturovaných informací z Wikipedie za účelem jejich zpřístupnění na webu pod licencemi Creative Commons Uveďte autora/Zachovejte licenci 3.0 Unported nebo GNU Free Documentation License. Články na Wikipedii tvoří většinou volný text, ale zahrnují také strukturované informace např. tzv. infoboxy, v nichž se prezentují základní fakta jednotlivých článků použitím standardizované šablony. Znalostní báze DBpedie je multijazyčná a v současnosti se skládá z více než miliardy RDF tripletů, které byly extrahovány z 92 jazykových verzí Wikipedie např. z anglické, německé, francouzské, španělské, italské, ale i české verze Wikipedie. Popisuje přes 3,4 milionů entit zahrnující osoby, místa, společnosti, nemoci apod. Pro každou z těchto entit DBpedie definuje globálně unikátní identifikátor. Dále zahrnuje 3,1 miliónů odkazů na externí webové stránky a 4,9 miliónů RDF odkazů na další webové datové zdroje.

ZÁVĚR

Sémantický web je koncept, k jehož realizaci a zároveň k obecnému zlepšení vyhledatelnosti elektronických zdrojů mohou knihovny a knihovníci přispět tvorbou různých slovníků a jejich implementací do formátů umožňujících vnoření webových technologií (např. užití jednotných identifikátorů URI, jazyků RDF, RDFS, OWL a jejich serializace v RDF/XML). Takové slovníky, tezaury a klasifikační schémata jsou v knihovnách vytvářeny už celá desetiletí, zatímco komunita sémantického webu je v této oblasti nováčkem. Spolupráce mezi knihovnami, komunitou sémantického webu a různými metadatovými iniciativami, jejichž společným cílem je pojmenování konceptů a entit a jejich vzájemné propojování, je klíčovým faktorem nejen v rozvoji systémů organizace znalostí, ale i sémantických webových technologií.

POUŽITÉ ZDROJE

1. *BIZER, Christian , et al. DBpedia : A crystallization point for the Web of Data. Web Semantics : Science, Services and Agents on the World Wide Web. 2009, vol. 7, issue 3, s. 154-165. ISSN 1570-8268.*
2. *BRAUN, Simone; ZACHARIAS, Valentin; HAPPEL, Hans-Jörg. Social Semantic Bookmarking [online]. 2008 [cit. 2010-04-29]. Dostupné z WWW: <http://www.slideshare.net/vzach/social-semantic-bookmarking-pakm-presentation>.*
3. *Common Tag [online]. 2009 [cit. 2010-03-11]. Dostupný z WWW: <http://www.commontag.org/Home>.*
4. *DBpedia [online]. 2009 [cit. 2010-04-30]. Dostupné z WWW: <http://dbpedia.org/About>.*
5. *Dublin Core : Czech homepage [online]. 2006 [cit. 2010-03-11]. Dostupný z WWW: <http://www.ics.muni.cz/dublin_core/index.html>.*
6. *JONES, Steve; et al. A transaction log analysis of a digital library. International Journal on Digital Libraries. 2000, vol. 3, no. 2, s. 152-169.*
7. *KOŽUCHOVÁ, Kristýna; ŠKUTA, Ctibor. Vliv trendů systémů organizace znalostí na vývoj Polytematického strukturovaného hesláře v Národní technické knihovně . Knihovna plus [online]. 2010, č. 1 [cit. 2010-04-01]. Dostupný z WWW: <http://knihovna.nkp.cz/knihovnaplus101/skuta.htm>. ISSN 1801-5948.*

8. *MYNARZ, Jindřich. Jak lze prakticky využít Polytematický strukturovaný heslář pro věcný popis elektronických zdrojů. Ikaros [online]. 2009, roč. 13, č. 12 [cit. 2010-03-01]. Dostupný na WWW: <http://www.ikaros.cz/node/5872>. URN-NBN:cz-ik5591. ISSN 1212-5075.*

9. *MYNARZ, Jindřich; KOŽUCHOVÁ, Kristýna; KAMRÁDKOVÁ, Kateřina. Novinky z oblasti Polytematického strukturovaného hesláře. Ikaros [online]. 2009, roč. 13, č. 7 [cit. 2010-03-04]. Dostupný na WWW: <http://www.ikaros.cz/node/5591>. URN-NBN:cz-ik5591. ISSN 1212-5075.*

10. *NĚMEČKOVÁ, Lenka; PAVLÁSKOVÁ, Eliška. Aplikace folksonomií v uživatelském rozhraní Jednotné informační brány. In Automatizace knihovnických procesů – 11 : sborník z 11. ročníku semináře pořádaného ve dnech 16.–17. května 2007 v Liberci. Praha : ČVUT, 2007. Dostupné z WWW: <http://www.akvs.cz/akp-2007/05-nemeckova-pavlaskova.pdf>. ISBN 978-80-01-03691-4.*

11. *PASTOR-SANCHEZ, Juan-Antonio; MARTINEZ MENDEZ, Francisco Javier; RODRÍGUEZ-MUÑOZ, José Vicente. 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, [cit. 2010-04-30]. Dostupný z WWW: <http://informationr.net/ir/14-4/paper422.html#authors>. ISSN 1368-1613.*

12. *RYLICH, Jan. Informační alchymie. Čtenář. 2009, roč. 61, č. 4, s. 150-152. Dostupný také z WWW: <http://ctenar.svkkl.cz/clanky/2009-roc-61/04-2009/knihovny-a-web-2-0-informacni-alchymie-57-382.htm>. ISSN 0011-2321.*

13. *SKLENÁK, Vilém. Metadata, sémantika a sémantický web. In INFORUM 2004 : 10. ročník konference o profesionálních informačních zdrojích, Praha 25.-27. května 2004. Praha : Albertina icome Praha, 2004 [cit. 2010-04-30]. Dostupné z WWW: <http://www.inforum.cz/pdf/2004/Sklenak_Vilem1.pdf>. ISSN 1801-2213.*

14. *SMRT, Martin. Proč analyzovat logy. Dobrý tip [online]. 2007 [cit. 2010-04-30]. Dostupný z WWW: <http://www.dobryweb.cz/newsletter-proc-analyzovat-logy/>.*

15. *SUMMERS, Ed, et al. LCSH, SKOS and Linked Data. In DC 2008 : International Conference on Dublin Core and Metadata Applications, Berlin 22.- 26. September 2008. Berlin : DCMI, 2008 [cit. 2010-04-30]. Dostupné z WWW: <http://dcpapers.dublincore.org/ojs/pubs/article/view/916/912>.*