ŠTÍTKOVÁNÍ JAKO PERSPEKTIVNÍ ALTERNATIVA K TAXONOMICKÉ KATEGORIZACI INFORMACÍ VE WEBOVÉM PROSTORU

Karel Michálek, Jana Filipová
Univerzita Pardubice, Fakulta ekonomicko-správní, Ústav systémového inženýrství a informatiky

Abstract: *The paper presents tagging as a perspective alternative to taxonomic categorization. The paper also deals with folksonomy, its advantages and disadvantage, s and the differences between folksonomy and taxonomy. It also mentions tagclouds as an alternative possibility in simplifying user interface. The end of this paper belongs to research of usage of tagging on Czech servers.*

Key words: *Tag, tagging, folksonomy, taxonomy*

1. Úvod

Pokud se chceme zabývat pojmem štítkování, je nutné jej zasadit do kontextu s informační architekturou, kterou jako první použil Američan Richard Saul Wurman již v roce 1976 [1,2]. Jenž začal uvažovat o souvislosti architektury s uspořádáním, organizováním a prezentací informací. Stejně jako architekt staví budovu podle potřeb jejích obyvatel, tak by i tvůrce jakéhokoliv informačního zdroje měl respektovat požadavky jeho uživatelů. Informační architektura se dá vymezit následovně:

1. jako věda a umění, která se zabývá organizováním informací;
2. představuje organizaci digitálních informací převážně v prostředí webových stránek a intranetů;
3. je chápaná úzce jako použití určitých metod a prvků na webových stránkách nebo je zaměňována s jinými obory.

V posledních letech se začíná rodit nový pojem, který úzce souvisí s informační architekturou a do jisté míry mění pohled na kategorizaci informací. Štítek (v některé české literatuře značka či návěští, což je dosti nejednoznačný překlad anglického slova tag) je nehierarchické klíčové slovo, které je využíváno k označení části informace. Informace může být reprezentována například digitálním obrázkem, textem, videem či jeho částí. Sloveso štítkování (tagging) je popularizováno až s pojmem Web 2.0 a stává se velice podstatnou vlastností velké části Web 2.0 služeb nebo nativních aplikací [3].

Štítkování je používáno převážně k zjednodušení vyhledávání a zpřehlednění uživatelského rozhraní. Štítkování značně zjednodušuje správu velkých datových souborů (internetových odkazů, produktů v internetových obchodech, článků v magazínech). Dále pak jsou štítky využívány k popisu objektů, nad kterými lze těžko provádět fulltextové vyhledávání např. rozsáhle databanky obrázků nebo videí. V tomto ohledu mohou být štítky chápány jako metainformace.

Štítkování je jeden z projektů v rámci konceptu Web 2.0, který dal možnost uživatelům širší participace na procesu tvorby, rozšiřování a vyhledávání informací. Ve vývoji předcházely štítkovacím systémům struktury vytvářené formou taxonomií, a to v souvislosti s automatizovanými mechanismy organizace internetových zdrojů. Odvození významu z textu pracuje na principu formálních výpočtů a je založeno více na přesnosti než na intuitivnosti či

zohledňování kontextu. Odezvou pak byl vznik sémantického webu [4], novějších značkovacích jazyků [5], mikroformátů, speciálních ontologií a metadatových schémat, která mají kódovat informace tak, aby byly lépe zpracovatelné stroji. Samotná skupina výzkumníků budující technologie sémantického webu si začala uvědomovat, že sémantický web je jen akademickou iluzí, protože se stroje v prostoru internetu nikdy nenaučí pracovat s informacemi tak, jak to dokáží lidé. Jiný přístup představuje kategorizace zdrojů na internetu lidmi. Příkladem je Dmoz: Open Directory Project [6], který je výstupem kolaborativní tvorby kategorií. Nové milénium odstartovalo vznik aplikací s otevřeným kódem a v souvislosti s nimi jsou vytvářené i struktury folksonomií. V roce 2003 Jozue Schacter přišel s projektem del.icio.us, jehož součástí byly a jsou „sociální odkazy“ (social bookmarks). Del.icio.us [7] patří mezi nejznámější a nejcitovanější systémy využívající folksonomii a štítkování, umožňuje správu a sdílení „oblíbených“ webových stránek.

Nedlouho poté, co byl spuštěn del.icio.us, se objevuje Flickr [8] umožňující sociální sdílení fotografií s využitím štítkování a to již od začátku svého vývoje. Českým zástupcem, který umožňuje sociální sdílení fotografií, je např. rajce.net.

2. Problematická taxonomie a rozvolněná folksonomie

Taxonomie je věda o klasifikaci. Taxonomické systémy jsou složeny z taxonomických jednotek, které jsou známé jako taxony a jsou uspořádány do hierarchické struktury [9]. Typicky je to závislost mezi subtypem a supertypem, tedy vztahu, nazývaného také vztah „rodič-dítě“ [10]. Typickým taxonomickým tříděním je např. květina je typu rostlina. Takže každá květina je také rostlina, ale ne každá rostlina je květina. Tedy taxonomie jsou svázané hierarchické struktury, které jsou předem definovány „klasifikační autoritou“. Schéma autority jako tvůrce taxonomie a uživatele jako příjemce informačního objektu je zachyceno na obr.1.

Obr.1: Proces zařizování a využívání informací v taxonomiích [zdroj: vlastní]

Příkladem může být zatřízení fotografií v adresáři, kde taxonomie nabízí tyto možnosti:
\fotografie\rodina\ - fotografie rodiny
\fotografie\dovolena\ - fotografie z dovolené
\fotografie\rodina\dovolena\ - fotografie, na kterých je rodina na dovolené
\fotografie\dovolena\rodina\ - fotografie z dovolené, na kterých je rodina

Zařazení jednotlivých fotografií do jednotlivých taxonů je relativně zásadním rozhodnutím vzhledem k pozdějšímu vyhledávání jednotlivých fotografií. Jednotlivé taxonomie mají rozdílnou sémantiku.

Od svého vzniku je celý web chápán jako taxonomická struktura. To dokazuje i systém URL adresace a systém záznamů na doménových serverech (DNS). Taxonomie je u domén dána doménami jednotlivých řádů. Ty jsou z pohledu taxonomie chápany jako taxonomy. Nejnadřazenějším taxonomem jsou národní domény (.cz, .com, apod.), dále jsou to taxonomy druhého řádu, doménová jména (upce.cz, uhk.cz, apod.). Následují pak domény (taxiomy) třetího řádu nebo také adresářová struktura webu (student.upce.cz, upce.cz/student), kde je

naráženo na stejnou problematiku, jako bylo výše naznačeno s problémem zatřízení fotografií do adresářů. Tento taxonomický pohled je posléze zaváděn i do pohledu na strukturování webového obsahu. Tedy veškeré informace, které jsou publikovány na webu, jsou taxonomicky tříděny do jednotlivých kategorií. S tím se setkáváme téměř napříč celým webem. Tato taxonomická struktura se odráží i v informační architektuře webové prezentace. Uveďme příklady:

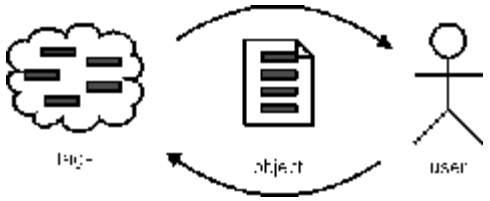
- *Elektronický obchod* – U elektronických obchodů je taxonomické třídění u jednotlivých druhů zboží, kde taxonomický strom popisuje vlastnosti zboží a na konci stromu se objevuje až konečný produkt (Nářadí → Profesionální nářadí → Vrtačky → Příklepová vrtačka).
- *Magazín* – Zde je taxonomie tvořena kategorizováním témat článků (Články → Domácí zpravodajství → Politika).

Novým typem kategorizace obsahu používaného výhradně na webu, který je vymezován převážně ve vztahu k rozvíjejícímu se Webu 2.0, je uživatelské třídění neboli folksonomie. Anglický pojem „folksonomy" odvodil roku 2004 informační architekt Thomas Vander Wal [11] od slov „folk" (lidé) a „taxonómie". Na rozdíl od centrálně řízených taxonomií zde sami uživatelé informací rozhodují a organizují, jakým způsobem budou informace třízeny. Toto třízení je prováděno pomocí štítků. Obecně je folksonomie využívána při práci s velmi rozsáhlými bázemi dat, které by bylo velmi náročné až nemožné zpracovat centralizovaně. V současnosti je spjata s novými typy aplikací, které nabízejí kolaborativní tvorbu obsahu. Výhody a nevýhody využití folksonomie shrnuje tab. 1. [12]

Tab. 1: Výhody a nevýhody folksonomie

Výhody folksonomie	**Nevýhody folksonomie**
• Využívá slovník tvořený uživateli (není nutný překlad do jazyka systému). • Uživatel může využít intuici. • Přispívá k budování komunit. • Je možné v krátké době přidat nové heslo. • V průběhu prohlížení uživatel objevuje nové informace. • Laciná alternativa klasických vyhledávacích systémů. • Hlavním akcentem je zde komunikace a sdílení. • Kontrolní nástroj pro hodnocení stávajících systémů.	• Mnohoznačnost. • Nepostihuje kontrolu synonym, homonym. • Vztahy jsou jen jednoúrovňové. • Nevhodné v případě rychlého přesného vyhledávání. • Nedostatek ochrany před neetickými uživateli. • Nutné určité množství uživatelů, aby byl systém důvěryhodný. • Nepoužívání standardů.

Zásadními rozdíly mezi taxonomickým a folksonomickým tříděním informací se zabývá Scott Golder a Bernardo Huberman [9] ve své práci, která jednoznačně dokazuje, že struktury vytvářené uživateli jsou daleko výhodnější. U folksonomie nejsou štítky tvořeny pouze experty, ale tvůrci či spotřebiteli jednotlivých obsahů (knih, článků, obrázků,…), kteří jsou známí, a je možné sledovat další jejich štítky. Schéma tvorby štítků je zachyceno na obr.2, kde je zřejmé, že tvůrcem i uživatelem štítků je uživatel a nikoliv autorita, jak je tomu v případě taxonomie viz obr.1.



Obr.2: Proces zařizování a využívání informací pomocí štítků [zdroj: vlastní]

V porovnání taxonomie a folksonomie jsou viditelné jasné rozdíly. Každý dokument může mít mnoho souvisejících termínů. Taxonomie jasně uvádí jednu klasifikaci pro jednu položku, má velmi hierarchické uspořádání a jasné vztahy. Folksonomie nemá hierarchickou strukturu a nejsou dány vztahy mezi termíny.

V současnosti dochází k syntéze folksonomie s tradičními nástroji na vyhledávání informací ve webovém prostoru [13]. V případě, že štítky budou výrazněji strukturované, mohly by být v budoucnu využity jako podklad pro vytváření pseudotaxonomického třízení, kde by mohla být odstraněna většina zmiňovaných nevýhod folksonomie.

V budoucnu může folksonomie sloužit pro konstrukci inteligentních agentů, kteří by se mohli učit od běžných uživatelů jak vytvářet štítky. Vznikaly by tak nové nástroje, které by využívaly mechanizmus organizace informací, ten by byl založen na umělé a výpočetní inteligenci a shlukovacích algoritmech (K-means, fuzzy shlukování, neuronové sítě, Kohonenovy samoorganizující se mapy apod.). Inteligentní agenti by potom mohli simulovat některé aspekty kategorizace webových objektů.

3. Problematická práce s databázemi při taxonomických strukturách

Taxonomicky kategorizovaná data jsou problematicky zpracovaná nejen uživateli, ale také v rámci relačních databází. V objektové databázi mohou být stromová data uložena přímo v takové podobě, jakou využívá aplikace, která se k této databázi připojuje. Naopak při použití relační databáze musí být data transformována tak, aby umožňovala uložení do ploché relační tabulky. Při čtení dat z databáze musí být zpětně transformována do podoby stromu. [14]

Nejznámějším a také nejčastěji využívaným způsobem, který lze při ukládání taxonomických struktur do relační databáze použít, je model, kde je součástí každého taxiomu také reference na rodičovský prvek. Nejvýše postavený prvek stromu, zvaný kořen, má referenci nastavenou na NULL. Pro získávání dat z takovéto tabulky se dá s úspěchem využít rekurzivní funkce. Pro zvýšení efektivity modelu může být datová struktura rozšířena o další atributy, které umožní rychlejší přístup k datům. Bude to atribut ORD (pořadí), který představuje pořadí uzlu v daném stromu, a atribut LEVEL, který představuje zanoření, respektive úroveň taxiomu.

Další možností je Modified Preorder Tree Traversal Algoritmus [15], kterým může být rozšířena datová struktura pro uložení taxonomické struktury. Princip spočívá v ohodnocení uzlů stromu dvěma hodnotami tím způsobem, že od kořene obcházíme všechny větve stromu a postupně se doplňuje pravá a levá hodnota uzlu, dokud se algoritmus nevrátí zpět ke kořenu. Kořen tím získává nejmenší levou a největší pravou hodnotu ze všech uzlů stromu. Všechny uzly nacházející se pod daným uzlem se dají získat dotazem na všechny uzly s levou hodnotou v intervalu pravé a levé hodnoty daného uzlu.

Ukládání štítků do relační databáze je velice jednoduché, ze své podstaty, kdy je není třeba transformovat do ploché struktury. Tedy každý záznam v databázi je obohacen o vazby na

jednotlivé štítky. Tímto jsou jednoduše obejity algoritmy spojené s ukládáním taxonomií. U objektových databází je toto ještě výrazně jednodušší.

4. Základní koncept štítkování

Z uživatelského hlediska je značkování činnost, při které je zdroji (obsahu webové prezentace) přiřazen jeden nebo více štítků. Tento štítek je přiřazen na základě uživatelské zkušenosti a obsahu zdroje, toto může být formalizováno:

Štítkování □ *{R, T_1...T_n};* kde R je zdroj a T_n jsou štítky přiřazeny ke zdroji *R*.

Pokud vezmeme v úvahu, že štítkování vzniká v sociálním prostoru a je vytvářeno uživateli v rámci folksonomie, je nutné formální zápis doplnit následujícím způsobem:

Štítkování □ *{R, T_1...T_n, TA_1...TA_i};* kde *TA* je autorem štítku.

Dále je štítkování nutné rozšířit o další parametr *S*, který zaručí obecnost štítkování. Parametr *S* definuje, ze kterého zdroje jsou štítky použity a to z důvodu, že kolekce (univerzum) štítků nemusí pocházet jen z jedné domény - webové aplikace. Obecný zápis štítkování je tedy zapsán následovně:

Štítkování □ *{R, T_1...T_n, TA_1...TA_i, S_1...S_j}.*

S takto formalizovaným zápisem lze dále pracovat při odstranění redundancí *T*.

Nabízí se myšlenka, jak využít štítkování v rámci sémantického webu (RDF), respektive ontologie. Tedy schematicky odlišit jednotlivé komponenty štítkování, tedy prvky množiny *{R, T_1...T_n, TA_1...TA_i, S_1...S_j}* tak, že jednotlivé prvky množiny nemají stejnou sémantickou váhu. Zde může být v úvahu brán např. model FOAF (The Friend of a Friend) pro identifikaci autora štítku *TA*, URI (Uniform Resource Identifier) pro zdroj štítků a ontologický slovník pro konkrétní štítek.

5. Zjednodušení uživatelských rozhraní pomocí štítkování „Tagclouds“

Tagclouds je stále oblíbenější aplikace štítků, jako druh navigace na webových stránkách (využívající technologie Web 2.0). Tento navigační prvek je tvořen ze všech štítků použitých na doméně, kde jednotlivé štítky jsou vizualizovány dle jejich popularity. Čím populárnější štítek, tím je v tagclouds výraznější viz obr.3.

Obr.3: Ukázka tagclouds [zdroj: www.smashingmagazine.com]

Dle [16] jsou tagclouds (*TC*) definovány jako množina *TC = (R, L)*, kde *R* □ *U* není prázdná množina zdrojů obsahující obecný prostor zdrojů *U*, nazývaný též universem. *L={(r, RID(p))* | *r*□*R, p*□*U}* je množina odkazů a *RID(p): R→A* je řídící funkce, která propojuje zdroje a adresy *A*.

Z předchozí definice si lze povšimnout, že se zavádí čistě abstraktní kategorie zdrojů, jako členů množiny zdrojů *R* prostřednictvím funkce *RID*. Není definováno, co je zdrojem, zda je

to slovo, událost nebo cokoliv jiného. Není zde tedy ještě definován sémantický význam zdroje. Zdrojem tedy může být cokoliv, co je definováno v *TC* jako odkaz. Odkaz *A* je tedy chápán jako podmnožina přirozeného jazyka interpretovaného v *TC*.

Tagclouds mohou být tedy reprezentovány jako distribuovaný systém znalostí dané domény, tím způsobem, že jsou prezentovány v přirozeném jazyce. Tedy pouze v tom případě, že je do *TC* připojen alespoň jeden zdroj. Jinak řečeno, *TC* jsou silným vizuálním navigačním prvkem, který reprezentuje znalosti dané webové prezentace na jednom místě pomocí odkazů na jednotlivé zdroje, které jsou na ní umístěny. Význam jejich využívání v pragmatickém hledisku je ve dvou rovinách a to pro vytváření vnitřních odkazů v dané prezentaci, a dále pak slouží především jako uživatelský filtr jednotlivých zdrojů (článků, odkazů, obrázků, apod.).

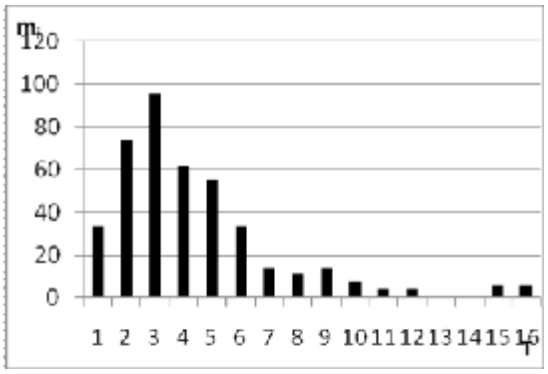
6. Využití štítkování na českém internetu

Pro zmapování využívání štítků na českých serverech bylo provedeno šetření, které proběhlo desátý měsíc roku 2008. Šetření se týkalo cca 200 domén, které mají vysokou návštěvnost (dle Navrcholu.cz), z toho bylo nalezeno pouze 19 domén, které využívají systém štítkování informačních objektů. Zjištěné domény byly dále rozděleny do čtyř hlavních kategorií.

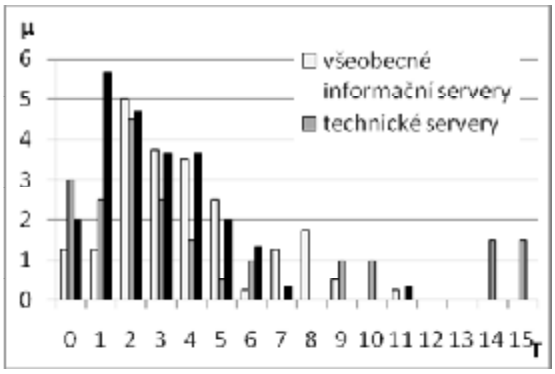
První kategorie nazvaná Všeobecné informační servery obsahuje osm domén. V této kategorii bylo zjištěno, že žádný z českých zpravodajských serverů (iDnes, iHned, Novinky) nevyužívá štítkování. Druhá kategorie byla nazvána Technologické servery, kde převažují zejména mobilní technologie. Celkem jsou v této kategorii 4 domény. Třetí kategorie byla nazvána Informatické servery a čítá 6 domén. Tato kategorie se věnuje internetu a informačním technologiím.

Poslední kategorii tvoří jediný server, a tím je server Českých Budějovic a Jihočeského kraje apu.cb.cz. Bohužel u žádného z ministerstev a zbylých krajů nebyla tato technologie nalezena. Při sledování využití technologie štítkování na českých vysokých školách bylo zjištěno, že tuto technologii využívá Univerzita Karlova a to v rámci IS FHS a Masarykova univerzita v rámci IS MU. Přehled serverů, které používají štítky, je umístěn v tabulce. V tabulce je zobrazena kategorie, do které byl server zařazen, dále název serveru, a pak četnost výskytu štítů na článek (aktualitu, stránku).

Absolutní četnost m_i je zobrazena v histogramu na obr.4. Nejčastěji se pro označení článků používá označení třemi štítky (T). Na obr.5 je pak zobrazen průměrný počet článků μ na počet štítků T, a to dle zaměření serveru (byla vyloučena kategorie Servery veřejné správy, protože obsahuje pouze jeden server a to není dostatečně reprezentující).



Obr.4: Proces zařizování a využívání informací pomocí štítků [zdroj: vlastní]



Obr.5: Proces zařizování a využívání informací pomocí štítků [zdroj: vlastní]

7. Závěr

Tento článek představuje štítkování jako perspektivní alternativu ke klasickým taxonomiím na internetu. V úvodní části byl definován proces štítkování a následně bylo provedeno šetření na českém internetu s akcentem na státní instituce. Z šetření vyplývá, že jen necelých 10 % nejnavštěvovanějších domén na českém internetu využívá štítkování a folksonomie jako metodu informační architektury pro zjednodušení orientace na internetových stránkách.

Domníváme se, že štítkování a foksonomie lze využívat jako plnohodnotné alternativy k taxonomiím. Jejich využití tak může výrazně zjednodušit práci s rozsáhlými datovými strukturami informačních objektů, především v rámci internetových, ale také intranetových aplikací.

Použitá literatura:

[1] WURMAN, R. S. Information Architects. [s.l.] : Watson-Guptill Pubns, 1997. 235 s. ISBN 978-1888001389.

[2] MAKULOVÁ, S. Informačná architektúra. Ikaros [online]. 2005, roč. 9, č. 9 [cit. 2008-05-04]. Dostupný z WWW: http://www.ikaros.cz/node/2007. URN-NBN:cz-ik2007. ISSN 1212-5075.

[3] O'REILLY, T. What Is Web 2.0 : Design Patterns and Business Models for the Next Generation of Software. O'Reilly [online]. 2005 [cit. 2008-10-20], s. 1-5. Dostupný z WWW: <http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html?page=1>.

[4] W3C Semantic Web Activity [online]. 2008 , 2008/10/18 [cit. 2008-10-20]. Dostupný z WWW: <http://www.w3.org/2001/sw/>.

[5] XHTML2 Working Group Home Page [online]. 2008 [cit. 2008-10-20]. Dostupný z WWW: <http://www.w3.org/MarkUp/>.

[6] DMOZ : Open Directory Project [online]. 2008 [cit. 2008-10-20]. Dostupný z WWW: <http://www.dmoz.org/>.

[7] Delicious : Social Bookmarking [online]. 2008 [cit. 2008-10-20]. Dostupný z WWW: <http://delicious.com/>.

[8] Flickr: Share your photos [online]. 2008 [cit. 2008-10-20]. Dostupný z WWW: <http://www.flickr.com/>.

[9] GOLDER, S. A., HUBERMAN, B. A. The Structure of Collaborative Tagging Systems. Journal of Information Science[online]. 2006 [cit. 2008-10-20], s. 198-208. Dostupný z WWW: <http://www.hpl.hp.com/research/idl/papers/tags/tags.pdf>.

[10] MAYR, E. The growth of biological thought: Diversity, evolution, and inheritance. Cambridge, MA: Harvard University Press. 1982

[11] WAL , Thomas Vander. Off the Top: Folksonomy Entries. Vanderwal.net [online]. 2008 [cit. 2008-10-20]. Dostupný z WWW: <http://www.vanderwal.net/random/category.php?cat=153>.

[12] HOLÁSEK, Daniel. Výhody a nevýhody folksonomií. Inflow: information journal [online]. 2008, roč. 1, č. 3 [cit. 2008-10-21]. Dostupný z WWW: <http://www.inflow.cz/vyhody-nevyhody-folksonomii>. ISSN 1802-9736.

[13] GREŠKOVÁ, M. Folksonómie v kontexte organizácie a vyhľadávania informácií [online]. Tlib. Informačné technológie a knižnice. 2006, č. 3 [cit. 2008-10-21]. Dostupný z WWW: <http://www.cvtisr.sk/itlib/itlib063/greskova.htm>. ISSN 1336-0779.

[14] ZELENKA, P. Metody ukládání stromových dat v relačních databázích. Interval.cz [online]. 2005 [cit. 2008-10-21]. Dostupný z WWW: <http://interval.cz/clanky/metody-ukladani-stromovych-dat-v-relacnich-databazich/>. ISSN ISSN 1212-8.

[15] VAN TULDER, G. Storing Hierarchical Data in a Database. SitePoint [online]. 2003 [cit. 2008-10-23]. Dostupný z WWW: <http://www.sitepoint.com/article/hierarchical-data-database/>.
[16] TOŠIČ, M., MILIČEVIČ, V. The Semantics of Collaborative Tagging System [online]. 2006 [cit. 2008-10-23]. Dostupný z WWW: <http://www.semanticscripting.org/SFSW2006/Paper6.pdf>.

Kontaktní adresy:
Ing. Karel Michálek, DiS.
Ing. Jana Filipová
Ústav systémového inženýrství a informatiky
Fakulta ekonomicko-správní
Univerzita Pardubice
Studentská 84, 532 10 Pardubice
Email: michalek@informacni.org
jana.filipova@upce.cz