Kategorie na Wikipedii z pohledu informační vědy

(text k prezentaci)

Leden 2011

Lucie Sakastrová
Ústav informačních studií a knihovnictví
Filozofická fakulta, Univerzita Karlova v Praze

Text k prezentaci z Wikikonference (http://www.wikikonference.cz/) se zaměřuje na nejvýznamnější vlastnosti a možnosti využití **kategoriálního systému** (folksonomie) encyklopedie **Wikipedie** (http://www.wikipedia.org). V první části jsou uvedeny základní shody a rozdíly kategoriálního systému (http://cs.wikipedia.org/wiki/Kategorie:Kategorie) a Wikipedie jako celku (viz snímek 2).

Wikipedie a její kategoriální systém se liší především tím, že Wikipedie je **encyklopedie**, zatímco její kategoriální systém **folksonomie**, tj. systém kolektivního třídění informací nebo organizování obsahu [Folksonomy, 2010]. Wikipedie i její kategoriální systém jsou tedy vytvářeny **kolektivně** a fungují na **wiki** principu. To znamená, že kdokoliv se může podílet na tvorbě jejich obsahu a kdokoliv jej může editovat [Wiki, 2010]. Tyto obsahy jsou navíc **otevřené**. To znamená, že je lze, za podmínky dodržení příslušných licencí, prakticky volně kopírovat, využívat a dále distribuovat [*Wikipedie*, 2011].

O **mnohojazyčnosti** a **multikulturnosti** encyklopedie i jejího kategoriálního systému lze pochybovat, neboť se nejedná ani tak o mnohojazyčnou encyklopedii, jako spíše o více než 250 jednojazyčných encyklopedií, které jsou vzájemně propojeny prostřednictvím tzv. mezijazykových odkazů (http://cs.wikipedia.org/wiki/Wikipedie:Mezijazykové_odkazy). Nicméně jejich obsahy nejsou totožné. Existují verze, které mají např. jen několik desítek článků, zatímco anglická verze má již více než 3,5 miliony článků [Wikimedia Foundation, 2011]. Výsledky každé jazykové verze jsou tak individuální a do jisté míry determinovány jak použitým jazykem, tak jeho „silou“ a kulturou či kulturami s ním spojenými. Stejné charakteristiky platí i pro kategoriální systém. Nejedná se tedy o jeden mnohojazyčný kategoriální systém, ale o několik desítek kategoriálních systémů jednotlivých jazykových verzí, které mohou být vzájemně propojeny mezijazykovými odkazy, nicméně jejich výsledky nejsou z hlediska rozsahu, struktury, zatříděných obsahů apod. totožné (viz např. porovnání snímku 6 a 7).

Nástroje řízení kvality jsou u kategoriálního systému stejné jako u Wikipedie jako celku. Patří sem nástroje preventivní, mezi které patří např. nápověda, dokumentace nebo kategorizování uživatelů podle přidělených práv do předem vytvořených skupin (http://cs.wikipedia.org/wiki/Speciální:Seznam_uživatelských_práv). Mezi nástroje následné lze zařadit mechanismy jako blokování (http://cs.wikipedia.org/wiki/Wikipedie:Blokování), mazání apod. (http://cs.wikipedia.org/wiki/Wikipedie:Údržba), mezi ty hybridní pak např. služby „Poslední změny“ (http://cs.wikipedia.org/wiki/Wikipedie:Poslední_změny) nebo „Sledované stránky“ (dostupná pouze po přihlášení) aj.

Následuje výčet významných shod a rozdílů mezi kategoriálním systémem Wikipedie a systémy třídění informací vytvářených odborníky (viz snímek 3).

Nejpodstatnějším rozdílem mezi kategoriálním systémem Wikipedie a systémy třídění informací vytvářených odborníky je riziko **vandalismu**, které hrozí prakticky jen v případě Wikipedie. Další charakteristiky a rizika, jako je **chybovost**, **redundance** nebo **inkonsistence** pak hrozí jak kategoriálnímu systému Wikipedie, tak systémům třídění informací obecně. Lze pouze předpokládat, že u systémů třídění informací vytvářených odborníky bude chybovost, redundance i inkonsistence nižší z důvodu vyšší pravděpodobnosti homogennosti této skupiny (znalost příslušného oboru, zaškolení do způsobu třídění informací, závazné normativy a příručky apod.). Dále lze předpokládat, že redundance, která se bude objevovat v systémech třídění informací vytvářených odborníky, bude mít alespoň předpoklad nějakého účelu a smyslu, přičemž ve Wikipedii tomu tak vždy být nemusí.

Mezi další významné charakteristiky systémů třídění informací obecně dále patří **míra podrobnosti** (tj. hloubka hierarchie) systému. Obecně lze říci, že míra podrobnosti bude u systémů třídění informací vytvářených odborníky pravděpodobně nižší (tj. menší hloubka hierarchie), jinak by se systém stal nepřehledným a neudržitelným. U kategoriálního systému Wikipedie je míra podrobnosti nejen neomezená, ale na různých místech stromů rozkladu též velmi individuální. Existují dokonce i kategorie věnované jednotlivcům (viz např. kategorie „Michael Jackson“ na snímcích 6 a 7), což je v případě systémů třídění informací vytvářených odborníky, jejichž účel je ryze vědecký, prakticky nemyslitelné (např. v případě biologické taxonomie kategorizovat druh „homo sapiens sapiens“ i na konkrétní příslušníky druhu apod.).

Samozřejmě, ani kategoriální systém Wikipedie, ani jiný systém třídění informací, nejsou objektivní, jelikož objektivita neexistuje. Existuje pouze **intersubjektivita** vyjadřující míru shody názorů libovolného množství zainteresovaných jednotlivců. U Wikipedie je takových jednotlivců více než u systémů informací vytvářených odborníky, navíc jsou mnohem různorodější. Intersubjektivita je tak u Wikipedie dána principy kolektivní tvorby, v případě systémů vytvářených odborníky pak mírou shody v oblastech jako je znalost příslušného oboru a názory na způsoby třídění informací apod.

Vzhledem k mnohem vyššímu počtu uživatelů Wikipedie oproti odborníkům spravujícím odborné systémy třídění informací, lze předpokládat větší **rychlost aktualizace** u Wikipedie a větší **rychlost domluvy** mezi odborníky spravujícími odborné systémy třídění informací. S tím souvisí i riziko **nestability** u Wikipedie a naopak **zkostnatělosti** a zastaralosti systémů třídění informací vytvářených odborníky, které nejsou schopny na změny vědeckého i jiného poznání reagovat dostatečně rychle. S počtem uživatelů dále souvisí riziko **zaujatosti**, přičemž lze předpokládat, že u Wikipedie bude, vzhledem k výrazně vyššímu počtu i různorodosti jejích uživatelů, pravděpodobně nižší. Ale ani to nelze konstatovat s jistotou.

Otázka **vědeckosti** jednoho vs. druhého systému je pak podobná jako otázka objektivity. Žádný systém třídění informací tak není 100% vědecký (např. vlivem nových objevů a nových pohledů na tyto objevy vyžadujících přepracování dosavadního stavu apod.). Nicméně lze předpokládat, že u systémů třídění informací vytvářených odborníky bude míra vědeckosti obecně vyšší než u kategoriálního systému Wikipedie, přičemž nelze vyloučit, že některé části kategoriálního systému Wikipedie jsou spravovány odborníky a rovněž tak reflektují určitou míru vědeckosti.

Následují konkrétní příklady problémových otázek kategorizování obsahů v české a anglické verzi Wikipedie (viz snímky 4-8).

Na prvním příkladě (viz snímek 4, http://cs.wikipedia.org/wiki/Miroslav_Plzák) je jako první demonstrována **problematika pojmenovávání** jednotlivých kategorií. Jako podivné se jeví např. spojení „vědečtí spisovatelé“, ze kterého lze jen těžko usoudit, co je tím konkrétně míněno. Jedná se pouze o vědce publikující ve svém oboru (což je prakticky každý vědec),

nebo snad vědce, kteří kromě vědeckých výsledků zároveň publikovali i beletrii, nebo snad vědce-popularizátory publikující odbornou literaturu popularizačního charakteru?

Dalšími zajímavými kategoriemi jsou **kategorie roků narození a úmrtí** (viz snímek 4, 6 a 7). U odborných databází by se zatřiďování osobností (např. vědců) do kategorií podle roků narození a úmrtí mohlo jevit jako příliš podrobné, možná až kontroverzní, nicméně tato problematika je zpravidla řešena alespoň prostřednictvím uvádění biografických dat v záznamech autorit.

Spojování minulosti osobností s organizací StB je v české kultuře stále velmi **kontroverzní záležitostí**, a je-li někdo ve Wikipedii do takové kategorie zařazen (viz snímek 4), je velmi vhodné si to minimálně ověřit v nějakém autoritativnějším zdroji, přičemž ani to v některých případech samozřejmě nemusí být zárukou jistoty, že tomu tak skutečně bylo.

Další zajímavou kategorií je pak kategorie „Manželství“ (viz snímek 4) a zejména **důvody zatřiďování** osobností do této kategorie. Zatřiďují se tam snad všechny osobnosti, které jsou nebo byly ve svazku manželském, nebo snad osobnosti, které se v souvislosti s manželstvím (např. mnohočetnými manželskými svazky) nějakým způsobem proslavily, či spíše osobnosti, které se problematikou manželství zabývaly nebo zabývají odborně?

Zajímavý je i výčet článků zatříděných do kategorie „Manželství“ v české verzi (viz snímek 5, http://cs.wikipedia.org/wiki/Kategorie:Manželství). Kromě některých kuriózních článků (např. „Kurtizána“), je zajímavé si povšimnout **obsahové provázanosti** kategorie „Manželství“ s problematikou náboženství v české verzi Wikipedie, tj. kultuře, která je v současnosti považována za velmi ateistickou. Znamená to snad, že kategorii „Manželství“ založili a udržují nábožensky smýšlející uživatelé Wikipedie, nebo že byla struktura této kategorie při vytváření inspirována obdobnou kategorií v jiné jazykové a tím např. i více nábožensky orientované verzi Wikipedie? Nebo že se česká kultura i přes rostoucí míru ateismu stále neoprostila provázanosti některých společenských institucí s náboženstvím?

Jako kontroverzní se dále může jevit otázka **kategorizování osobností podle barvy pleti**, rasy či etnického původu (viz kategorie „Afroameričané“ na snímku 6, http://cs.wikipedia.org/wiki/Michael_Jackson). V odborné databázi by se kategorizování autorů nebo vědců podle barvy pleti pravděpodobně jevilo nejen jako kontroverzní, ale třeba i rasistické či xenofobní, a tudíž nepřijatelné.

Porovnáním kategorií na snímku 6 (http://cs.wikipedia.org/wiki/Michael_Jackson) a 7 (http://en.wikipedia.org/wiki/Michael_Jackson) lze dospět k závěru, že Michael Jackson byl pro anglo-americkou jazykovou **kulturu** významnější osobností než pro jazykovou kulturu českou. V anglické verzi Wikipedie je článek o Michaelu Jacksonovi zatříděn do více kategorií obsahově zacházejících do těch **nejdetailnějších podrobností**. Jiným důvodem těchto rozdílů pak samozřejmě může být i větší rozsah a detailnější propracovanost kategoriálního systému anglické verze oproti české.

Jako **kuriózní** se jeví i kategorie „Lidé podle stavu“ (viz snímek 8, http://cs.wikipedia.org/wiki/Kategorie:Lidé_podle_stavu), zvláště pak kategorie „Možná žijící lidé“ a „Neexistující lidé“, ale i ostatní v této kategorii. Jediné kategorie přijatelné i pro odborné databáze jsou kategorie „Žijící lidé“ a „Zemřelí lidé“, které jsou v těchto systémech vyjadřovány prostřednictvím uvádění biografických dat v záznamech autorů, vědců apod. Dalším přípustným hlediskem kategorizace autorů nebo vědců v odborných databázích je kategorizace podle předmětu odborného zájmu. Ostatní hlediska v kategorii „Lidé podle stavu“ by se pak pravděpodobně pro účely kategorizování autorů či vědců v odborných databázích jevila jako **nepřípustná**, nicméně ve Wikipedii má takováto kategorizace osobností (a tím i vědců) svůj smysl a účel (viz dále).

Závěrečné části prezentace pojednávají o rozdílech využití kategoriálního systému Wikipedie a systémů třídění informací vytvářených odborníky se zaměřením na význam kategoriálního systému Wikipedie pro účely budování sémantického webu (viz snímky 9-10).

Využívat kategoriální systém Wikipedie je tak vhodné spíše pro **účely zájmu**, **koníčku nebo zábavy**, zatímco systémy třídění informací vytvářené odborníky pro **účely odborné** (viz snímek 9). **Přesnost** a **úplnost** vyhledaných informací ve Wikipedii není tak důležitá jako přesnost a úplnost vyhledaných informací v odborných databázích, kde např. v případě patentové databáze nevyhledání byť jednoho jediného dokumentu může v případném průmyslově-právním sporu znamenat významnou finanční ztrátu. Zatímco systémy třídění informací vytvářené odborníky jsou zpravidla určené jak pro **prohlížení**, tak pro **vyhledávání** (samostatné vyhledávací pole, rotovaný rejstřík apod.), kategoriální systém Wikipedie se hodí spíše pro prohlížení, i když vyhledávat v něm lze také (abecední rejstřík kategorií, pokročilé vyhledání s omezením na obsah typu kategorií apod.).

Využívat systémy třídění informací vytvářené odborníky je tak vhodné v **odborných zdrojích tvořených a spravovaných profesionály**, zatímco kategoriální systém Wikipedie v **otevřených zdrojích tvořených kolektivně** a dále v externích systémech a službách pracujících s obsahem Wikipedie, jako jsou **internetové vyhledávače**, **linkovací a referenční služby** nebo **služby sémantického webu** (tj. služby nového evolučního stupně stávajícího webu, kde jsou informace strukturovány a uloženy podle standardizovaných pravidel, což usnadňuje jejich vyhledání a zpracování [Sémantický web, 2010]).

Zatímco v případě databází a vyhledávačů je zpravidla nutné vyhledávací dotaz formulovat prostřednictvím více či méně složitého **dotazovacího jazyka**, u služeb sémantického webu lze dotazy formulovat i ve větách **jazyka přirozeného** (viz snímek 10). Příslušná služba je schopná dotaz formulovaný v přirozeném jazyce převést do jazyka dotazovacího tak, aby s ním byl schopen pracovat stroj a na výstupu podat přijatelné výsledky.

Např. v případě dotazu „Vypiš mi všechny americké spisovatele 19. století“ by takový **stroj** mohl pracovat zhruba tak, že v příslušné jazykové verzi Wikipedie nejprve prozkoumá příslušnou existující kategorii (v české verzi např. nazvanou) „Spisovatelé podle národnosti“ (http://cs.wikipedia.org/wiki/Kategorie:Spisovatelé_podle_národnosti) a vybere podkategorii „Američtí spisovatelé“ (http://cs.wikipedia.org/wiki/Kategorie:Američtí_spisovatelé). Následně se podívá na kategorii „Spisovatelé podle století“, existuje-li, a vybere podkategorii „Spisovatelé 19. století“. Neexistuje-li, podívá se např. na kategorie s daty narození a úmrtí spisovatelů zařazených do kategorie „Američtí spisovatelé“ apod. Následně porovná články v podkategorii „Spisovatelé 19. století“ (nebo kategorie s daty úmrtí a narození spisovatelů v podkategorii „Američtí spisovatelé“) s články v podkategorii „Američtí spisovatelé“ a na výstupu podá soupis jen těch článků, které jsou zařazeny jak do kategorie „Spisovatelé 19. století“, tak do kategorie „Američtí spisovatelé“, nebo jen ty zařazené do kategorie „Američtí spisovatelé“, jejichž data úmrtí a narození odpovídají 19. století. U složitějších dotazů, jako je např. ten na snímku 10, je pak takových kroků samozřejmě více a porovnávání obsahů zařazovaných do souvisejících kategorií též složitější.

I když se tedy některé způsoby a výsledky kategorizování obsahu ve Wikipedii mohou z hlediska **informační vědy** jevit na první pohled jako náhodné, nadbytečné, kontroverzní, neúčelné nebo nesmyslné, ve službách **sémantického webu** a podobných moderních službách nacházejí zajímavé uplatnění, jehož **význam** má již v současnosti výraznou stoupající tendenci nejen v oboru informační vědy.

Literatura

- Folksonomy. In *Wikipedia : the free encyclopedia* [online]. San Francisco (Kalifornie, USA) : Wikimedia Foundation, 2001- , last modified on 16 December 2010 at 12:23 [cit. 2011-01-08]. Anglické rozhraní. Dostupné z WWW: <http://en.wikipedia.org/wiki/Folksonomy>.
- Sémantický web. In *Wikipedie : otevřená encyklopedie* [online]. San Francisco (Kalifornie, USA) : Wikimedia Foundation, 2002- , naposledy editována 26. 11. 2010 v 18:56 [cit. 2011-01-09]. České rozhraní. Dostupné z WWW: <http://cs.wikipedia.org/wiki/Sémantický_web>.
- Wiki. In REITZ, Joan M. *ODLIS : Online Dictionary of Library and Information Science* [online]. Westport (CT) : Libraries Unlimited, 2004-2010, last updated March 9, 2010 [cit. 2011-01-08]. Vyšel i v tištěné formě. Dostupné z WWW: <http://lu.com/odlis/>.
- Wikimedia Foundation. 2011. *Wikimedia : meta-wiki* [online]. San Francisco (Kalifornie, USA) : Wikimedia Foundation, 2001-, last modified on 4 January 2011, at 19:09 [cit. 2011-01-08]. List of Wikipedias. Anglické rozhraní. Dostupné z WWW: <http://meta.wikimedia.org/wiki/List_of_Wikipedias>.
- *Wikipedie : otevřená encyklopedie* [online]. San Francisco (Kalifornie, USA) : Wikimedia Foundation, 2002- , naposledy editována 3. 1. 2011 v 13:16 [cit. 2011-01-08]. Wikipedie. České rozhraní. Dostupné z WWW: <http://cs.wikipedia.org/wiki/Wikipedie>.