Nástroje Google.
3. Google Book Search

Miroslav Bartošek, ÚVT MU

Služba Google Book Search představuje snahu rozšířit osvědčené googlovské technologie z oblasti webu do dalšího informačního prostoru – do oblasti tištěných knih. Vzhledem k obrovskému množství existujících knih (některé odhady uvádí 100 miliónů titulů vytvořených od počátku psaných dějin lidstva [4]) jde o prostor informačně bohatý, kvalitní a pro uživatele nepochybně užitečný. Současně jde ale také o prostor komerčně velmi zajímavý. A právě komerční stránka věci spolu s otázkou autorských práv vyvolala u této služby největší polemiku a právní spory. Koncem loňského roku došlo k zásadní dohodě mezi firmou Google a hlavními odpůrci, která otevírá možnost další rozvoj služby výrazně akcelerovat.

1 Co to je

Google Book Search (GBS) `http://books.google.com` vyhledává knihy, a to na základě fulltextového prohledávání jejich obsahu (nikoliv jen prohledáváním bibliografických záznamů, jak to dělají běžné knihovní systémy). Jde o knihy jakéhokoliv typu – beletrii i knihy odborné. Protože texty většiny knih nejsou v elektronické podobě dostupné, je součástí GBS projekt masové digitalizace tištěných knih; výhledově Google plánuje digitalizovat 30 až 60 miliónů titulů! Skenování knih a zpřístupnění jejich obsahu nese s sebou ale jeden zásadní problém, a tím jsou autorská práva – copyright.

Na rozdíl od volně přístupného webu, kde problém autorských práv není z pohledu vyhledávacích služeb tak palčivý (i když nelze tvrdit, že neexistuje), je v případě knih situace diametrálně odlišná. Knihu obvykle autor resp. držitel práv nevystavil na web k volnému použití. Může tedy někdo, bez explicitního souhlasu autora, vůbec takovou knihu skenovat (převádět z tištěné do elektronické formy), indexovat a umožnit v ní komukoliv na webu vyhledávat? To bylo a je hlavním jádrem sporu.

Google řešil tento spor pragmaticky. Rozdělil knihy z pohledu autorských práv do tří skupin. V první skupině jsou knihy, u kterých již autorská ochrana vypršela[1]. Zde není zásadní problém – tyto knihy jsou ve veřejném vlastnictví (public-domain), takže Google je může bez problémů skenovat, umožnit v nich vyhledávat a dokonce může zpřístupnit bez omezení i plné texty takových knih (uživatel si je může číst na obrazovce počítače nebo stáhnout v podobě PDF-souboru). Druhou skupinu představovaly knihy chráněné copyrightem, k jejichž zařazení do GBS neměl Google explicitní souhlas držitele práv (současně však Google zveřejnil, že pokud vlastník nebude s digitalizací svého díla souhlasit, bude jeho nesouhlas respektovat a knihu skenovat nebude). Tyto knihy Google skenoval a umožnil v nich vyhledávat; nezobrazuje však již uživatelům plný text, pouze několik krátkých úryvků obsahujících okolí hledaného výrazu. Třetí skupinu představují knihy chráněné copyrightem, u nichž držitelé práv poskytli souhlas se zařazením do GBS. Google tyto knihy skenuje (nebo přebírá jejich elektronickou podobu) a držitel práv sám určí, jaká část knihy bude uživateli v GBS zobrazována.

Služba byla poprvé představena v roce 2004, tehdy ještě pod poněkud zavádějícím názvem Google Print. Podobně jako stejně starý Google Scholar je i Google Book Search dodnes označován jako „beta verze“ – nicméně služba se neustále zdokonaluje a pokroky jak v obsahu databáze (počty knih), tak i softwaru jsou velmi povzbudivé. Dle údajů [3] nabízela služba GBS koncem roku 2008 již přes 7 miliónů knih, z toho přes jeden milión tvořily knihy ve veřejném vlastnictví, tedy s volně dostupnými plnými texty. A tyto počty se rychle zvyšují.

Odkud vlastně Google všechny tyto knihy bere? Využívá k tomu dva programy, které jsou součástí GBS – Projekt knihovna (Library Project) a Partnerský program (Partner Program).

[1] Doba autorskoprávní ochrany díla (copyright) může být v různých zemích různě dlouhá. V zemích EU a USA je to od okamžiku vytvoření díla do 70 let po smrti autora; u „najatých děl" (ve vlastnictví korporací) je v USA ochrana 95 let od publikace díla.

Projekt knihovna zahájil v roce 2004 spoluprací s pěti významnými knihovnami s rozsáhlými knihovními fondy: knihovnami Michiganské univerzity, Harvardské univerzity, Stanfordské univerzity, Oxfordské univerzity a Newyorské veřejné knihovny. Postupně se zapojovaly další. Knihovny poskytují knihy, které Google skenuje speciálně k tomu vyvinutými super-výkonnými technologiemi a dále je zpracovává (rozpoznání textu pomocí OCR, indexace textu, doplnění základních bibliografických metadat, vazeb na jiné informační zdroje a dalších užitečných údajů – vše je zpracováváno automatizovaně). Jde přitom skutečně o velmi masivní produkci. Například jenom dle smlouvy s Michiganskou univerzitou má být během šesti let zpracováno 7 miliónů knih z fondů univerzitní knihovny (i při nepřetržitém provozu 24 hodin denně, 365 dnů v roce by to znamenalo zpracovat více jak dvě knihy každou minutu!).

Partnerský program je zaměřen na vydavatele a autory. Umožňuje jim, aby sami poskytli své knihy k zařazení do GBS (buď dodáním tištěných knih ke skenování nebo nahráním elektronické verze knih do databáze GBS). Za to jim Google nabízí lepší on-line marketing (zvýšení viditelnosti knih a také webových stránek vydavatelů), zvýšení prodeje (u autorsky chráněných knih neposkytuje GBS uživatelům plné texty, ale přesměrovává je na knihkupectví, kde si mohou danou knihu koupit), a podle nedávno uzavřené dohody dokonce i finanční podíl z příjmů na kontextové reklamě (nový zdroj zisků). V současnosti je do programu zařazeno již přes 20 000 partnerů.

2 Začínáme s Google Book Search

Domovská stránka služby Google Book Search `http://books.google.com` se nijak zvlášť neliší od toho, na co je uživatel zvyklý u webového vyhledávače Google nebo u Google Scholar. Základní vyhledávání nabízí googlovsky jednoduchou obrazovku s jediným polem pro zadání hledané fráze (na úvodní stránce anglické verze se zobrazuji ještě i obálky vybraných knih a seznamy knih podle oborů). Rozšířené vyhledávání pak nabídne formulář pro přesnější specifikaci – například lze specifikovat jméno autora, název knihy, jazyk, vydavatele, datum vydání či ISBN. Současně je možno nastavit vyhledávací filtr, který omezí vyhledávání podle dostupnosti textu (např. jen knihy s volně dostupnými plnými texty – s úplným zobrazením) nebo podle druhu dokumentu (lze vyhledávat nejen v knihách ale i v populárních časopisech – magazínech, které Google začal přidávat do GBS od prosince 2008; jako příklad takového časopisu uveďme Popular Science Magazine[2]).

Jako výsledek vyhledávání se zobrazí seznam relevantních knih. Kliknutím na zvolenou knihu přejde uživatel na *referenční stránku knihy*. Rozsah informací a služeb na referenční stránce závisí na tom, do které ze čtyř kategorií – z hlediska možností zobrazení textu – kniha patří (od nejjednodušší k nejbohatší):

Náhled není k dispozici (No preview available): nejrestriktivnější kategorie, kdy jsou uživateli o dané knize poskytnuta jen základní bibliografická data (obdoba zjednodušeného záznamu v lístkovém katalogu); žádná část textu knihy není přístupná. Příklad: `http://books.google.com/books?id=BOmbA`

Zobrazení fragmentů (Snippet view): kromě základních bibliografických údajů je uživateli zobrazeno i několik (nejvýše tři) fragmentů (snippets) – vět z knihy zobrazujících hledaný výraz v kontextu. Uživatel může v textu knihy dále vyhledávat a zobrazovat si jiné fragmenty (v omezeném množství). Příklad: `http://books.google.com/books?hl=cs&id=G3NLAAAAMAAJ&q=Franci&pgis=1`

Omezený náhled (Limited preview): v této kategorii je uživateli zobrazen omezený počet stran textu. Rozsah stanovuje držitel práv, obvykle bývá zobrazeno kolem 20 % stran příslušné knihy. Uživatel tak může danou knihou „listovat" obdobně, jako by si namátkově prohlížel fyzickou knihu v knihkupectví. Příklad: Kniha *The Calculus Galery* z roku 2005 `http://books.google.com/books?vid=ISBN691095655&hl=cs`.

Úplné zobrazení (Full view): informačně nejbohatší kategorie, kdy je uživateli k dispozici plný text celé knihy. Úplné zobrazení je možné v případě, kdy je kniha ve veřejném

[2] `http://books.google.com/books?id=Ok8XtrhowscC&hl=cs`

vlastnictví (nevztahují se již na ni autorská práva)[3] nebo když vydavatel či autor požádal, aby byla kniha plně viditelná. Úplné zobrazení umožňuje prohlédnout si kteroukoli stránku příslušné knihy, a pokud je kniha ve veřejném vlastnictví, lze si ji rovněž stáhnout ve formátu PDF[4]. Jako příklad uveďme Komenského *Orbis Pictus* z roku 1833 `http://books.google.com/books?id=9uoIAAAAQAAJ&hl=cs` či Euklidovy *Elements* `http://books.google.com/books?id=9ViEZbTGaeEC&hl=cs`.

Jsou-li k dispozici, mohou být u každé z výše uvedených kategorií knih poskytnuty další užitečné informace: obálka, obsah, oblíbené pasáže, další vydání, recenze, odkazy z webových stránek, odkazy z vědeckých prací, odkazy z knih, místa zmíněná v knize s vyznačením pozic na mapě Google-maps[5] aj.

Současně jsou na každé referenční stránce knihy umístěny odkazy na knihkupectví, kde si uživatel může knihu koupit, a na nejbližší knihovnu, kde se kniha dá vypůjčit (tato funkce je realizována odkazem do celosvětového katalogu WorldCat společnosti OCLC, který se po zadání země nebo kódu PSČ pokusí nalézt místní knihovnu vlastnící daný knižní titul).

Registrovaný uživatel si také může zřídit v rámci GBS vlastní knihovničku, psát recenze, přidělovat knihám štítky a u knih v úplném zobrazení dokonce anotovat části textu. Tyto své vlastní informace může pak sdílet s jinými uživateli ať již přímo (URL odkazy či RSS kanály), nebo v rámci dalších služeb Google, jako například Google Blogger či Google Notebook.

3 Dohoda s vydavateli – a co z ní vyplývá

Představitelé Google od počátku deklarovali, že kladou silný důraz na dodržování autorských práv. Přesto však projekt GBS narážel záhy po svém uvedení na zásadní odpor velkých vydavatelů. Nelíbilo se jim zejména to, že Google začal skenovat knihy pod autorskoprávní ochranou bez explicitního souhlasu držitele práv. I když Google neposkytoval u těchto knih plný text uživatelům a využíval ho pouze pro indexaci obsahu a vyhledávání, a přestože nabídl držitelům práv možnost opt-out – tj. určit knihy, které budou ze skenování vyloučeny, vydavatelé se cítili ohroženi a finančně poškozeni.

Spory vyvrcholily žalobou za rozsáhlé narušení autorských práv (massive copyright infringement), kterou na Google podaly v roce 2005 organizace Cech amerických autorů, Asociace amerických nakladatelů a další. Předpokládalo se, že se soudní spor povleče dlouhou řadu let. Proto bylo poněkud překvapivé, když bylo 28. října 2008 oznámeno, že mezi oběma stranami sporu byla uzavřena *dohoda o vyrovnání* (ta zatím nebyla soudně potvrzena, soud o ní bude jednat až ve druhé polovině roku 2009).

Dohoda je velmi obsáhlá a složitá – i s dodatky má přes 200 stran. Uznává práva a zájmy držitelů autorských práv, nabízí jim kontrolu nad tím, jak budou jejich knihy v GBS využívány a poskytuje jim podíl z příjmů Google za kontextovou reklamu. Současně Google uhradí soudní výlohy a v rámci odškodnění za již naskenované chráněné knihy investuje částku 125 miliónů dolaru do nové nezávislé neziskové organizace *Registr autorských práv*, která bude zastupovat autory, vydavatele i další vlastníky autorských práv. Organizace bude pomáhat vyhledávat vlastníky autorských práv a zajistí, aby tito získali peníze, které si na základě této dohody vydělají.

Díky dohodě mají být nabídnuty nové možnosti přístupu k plným textům autorským zákonem chráněných knih:

[3] Vzhledem k tomu, že v různých zemích platí různá pravidla pro autorskoprávní ochranu, nemusí být vůbec snadné určit, zda je daná kniha ve veřejném vlastnictví či nikoliv. V případě USA to v současnosti většinou znamená, že kniha musela být vydána před rokem 1923. V případě zemí mimo USA se Google řídí místními zákony, přičemž při interpretaci daného autorského zákona a známých faktů o konkrétní knize zachovává konzervativní přístup. Uživatelé mohou upozorňovat na knihy, které jsou ve veřejném vlastnictví, a přesto je Google nenabízí v úplném zobrazení.

[4] Vedle čtení ve formátu PDF existuje i možnost „Prohlížet jako prostý text“. Tato možnost otevírá knihu adaptivním technologiím, jako jsou například čtečka obrazovky či Braillův displej, a umožňuje tak lepší přístup uživatelům s vadou zraku.

[5] V aplikaci Google Earth je služba, která funguje přesně opačně – uživatel si vybere místo a Google mu řekne, které knihy s ním souvisí.

- Individuální přístup on-line: jednotliví uživatelé budou mít možnost zakoupit si on-line přístup k plným textům miliónů chráněných knih a přistupovat k nim přes svou osobní knihovničku (jako registrovaní uživatelé);
- Přístup pro knihovny a univerzity: knihovny, univerzity a další instituce budou mít možnost zakoupení přístupové licence pro celou organizaci.

Lepší dostupnost by se měla týkat zejména tzv. vyprodaných knih (out-of-print books), které doposud byly k sehnání pouze v knihovnách či antikvariátech. Nyní budou široce dostupné on-line přes GBS – a to bezplatně formou omezeného náhledu nebo v režimu úplného zobrazení za poplatek. Veřejné knihovny v USA budou také moci nabídnout jeden terminál v budově pro veřejný bezplatný přístup.

Nově se také otevírá příležitosti pro badatele, kteří budou moci využívat korpus miliónů knih v indexu GBS pro výzkumné účely.

Protože uvedená dohoda řeší soudní spor v USA, týká se přímo pouze uživatelů, kteří přistupují ke službě GBS v USA. Mimo území USA bude služba fungovat stejně jako doposud. Do budoucna se však zřejmě bude Google snažit dosáhnout obdobné dohody i se zahraničními vlastníky autorských práv.

Je třeba říci, že ne všichni tuto dohodu přivítali. Nespokojeni jsou zejména ti, kteří očekávali, že soud potvrdí jejich přesvědčení, že skenování knih za účelem jejich indexace a vyhledávání spadá pod tzv. *fair use* – tj. taková užití díla, na která se nevztahuje autorskoprávní ochrana. Z tohoto pohledu dohoda oslabuje pozici uživatelů a vývojářů informačních služeb vůči držitelům autorských práv. A navíc se mnozí obávají, že podobných „čertových kopýtek“ může dohoda skrývat více, včetně komercionalizace služby a nežádoucího posílení monopolního postavení Google. Mezi těmi, kdo v souvislosti s dohodou vyslovili velké znepokojení, byla i Harvardská univerzita – jedna z pěti prvních knihoven zapojených do knihovního projektu GBS (univerzita údajně dokonce z projektu odstoupila).

4 Další podobné projekty

Google Book Search není prvním ani jediným projektem v oblast masové digitalizace knih a jejich on-line zpřístupnění. Nejstarším z nich je *projekt Gutenberg* `http://www.gutenberg.org` zahájený již v roce 1971, s cílem digitalizovat s pomocí dobrovolníků především anglická klasická literární díla ve veřejném vlastnictví. V současnosti obsahuje kolem 27 000 knih dostupných v čistě textovém formátu (plain-ASCII).

Hlavním konkurentem projektu Google Book Search je v současnosti *Open Content Alliance* `http://www.opencontentalliance.org/`, kooperativní projekt založený v roce 2005 společnostmi Internet Archive a Yahoo! (postupně se zapojila řada dalších) s cílem masové digitalizace a trvalé archivace knih. Knihy pod autorskoprávní ochranou se od začátku skenují výhradně až po udělení souhlasu majitele autorských práv. Výstupem této iniciativy je volně dostupná digitální knihovna *Open Library* `http://openlibrary.org/`, která nabízí v současnosti téměř 23 miliónů záznamů knih (z toho jeden milión i s plnými texty).

Dalším konkurentem měl být projekt Microsoftu s názvem *Windows Live Book Search*, ten však byl v květnu 2008 zastaven.

Projekt Carnegie Mellon University s názvem *Universal Digital Library* (UDL) `http://www.ulib.org/` digitalizuje knihy od roku 2001 a dosáhl již více než 1,5 miliónu titulů (milión z nich v čínštině). Na projektu spolupracuje 50 skenovacích center po celém světě. V letech 2006-2007 byl v rámci aktivit UDL realizován projekt the Milion Book Digital Library, jehož cílem bylo ověřit technologie pro masovou digitalizaci.

V knihovním světě je nejznámějším a nejrozsáhlejším digitalizačním počinem projekt Kongresové knihovny USA s názvem *American Memory* `http://memory.loc.gov`, který od poloviny 90. let digitalizoval na deset miliónů informačních objektů z historických sbírek Kongresové knihovny a dalších amerických knihoven (knih, dopisů, dokumentů, fotografií, map, zvukových nahrávek, filmových záznamů atd.)

5 Závěr

Google Book Search je nepochybně užitečná a slibně se rozvíjející služba. I když Google uvádí, že „cílem je pomáhat uživatelům nacházet knihy a zjišťovat, kde se dají koupit nebo půjčit, nikoli číst je celé on-line“, nabízí i velké množství plných textů knih ve veřejném vlastnictví. Služba má samozřejmě i své stinné stránky. Patří mezi ně například ne vždy dostatečná kvalita skenů a OCR textů, převaha anglicky psané literatury (a z toho pramenící obavy jinojazyčných národů z „kulturního imperialismu“), obavy z monopolizace trhu a nejistota ohledně dopadů nedávné dohody mezi Googlem a vydavateli. Pro naše uživatele je zatím nevýhodou i malé pokrytí českých knih. Přesto však přínosy jednoznačně převažují. Hlavním přínosem je služba samotná, významné však jsou i dopady do oblasti digitalizace – nové výkonné technologie pro masovou digitalizaci a razantní pokles cen skenování.

Literatura

[1] O službě Vyhledávání knih Google. http://books.google.com/intl/cs/googlebooks/about.html

[2] Dian Schaffhauser. Google Book Search: The Good, the Bad, & the Ugly. Campus Technology. 1.1.2008. http://campustechnology.com/Articles/2008/01/Google-Book-Search-The-Good-the-Bad-amp-the-Ugly.aspx

[3] Wikipedia: Google Book Search. http://en.wikipedia.org/wiki/Google_Book_Search

[4] Universal Digital Library (UDL). http://www.ulib.org/ □