Nástroje a služby Google. 2. Google Scholar

Miroslav Bartošek, ÚVT MU

Téměř přesně před čtyřmi lety, v listopadu 2004, představila firma Google svůj nový produkt – vyhledávač vědecké literatury *Google Scholar* (GS). Jeho uvedení vzbudilo velkou pozornost a rychle se vynořila záplava článků, recenzí a natěšených uživatelů. Prvotní nadšení však brzy částečně ochladlo, když se ukázalo, že tento nástroj – jakkoliv pro řadu uživatelů nesporně velmi užitečný, má i mnoho zásadních vad na kráse. Vad, které navíc často přetrvávají až do dnešních dnů. Jak to tedy s Google Scholar vypadá? Je to spící tygr s potenciálem monopolizovat v budoucnu oblast vyhledávání vědeckých informací, obdobně jako klasický vyhledávač Google dominuje dnes obecnému vyhledávání na webu? Nebo je to ze strany Google spíše již „odpískaný" nedotažený pokus, slepá ulička vývoje? Těžko hádat, ponechme to budoucímu vývoji. Dnes si alespoň představme Google Scholar jako takový – jeho aktuální stav se všemi obecně uváděnými plusy a mínusy.

1 Co to je

Google Scholar, `http://scholar.google.com`, je služba pro vyhledávání digitálních i fyzických kopií *odborných a vědeckých prací:* recenzovaných časopiseckých a sborníkových článků, preprintů, technických zpráv, odborných knih, vědeckých kvalifikačních prací atd. Cílem je usnadnit uživateli vyhledání potřebné vědecké literatury a nezahrnout ho přitom neodborným balastem, kterým oplývá běžný web. Vyhledávány jsou nejen dokumenty volně dostupné na webu, ale i články v časopiseckých kolekcích akademických nakladatelů, práce v otevřených i uzavřených full-textových, abstraktových a bibliografických databázích, repozitářích učených společností a univerzit, katalozích odborných knihoven. Ne vše, co Google Scholar indexuje a nabízí, musí být nutně (volně) dostupné v plnotextové podobě. Kromě klasického vyhledávání nabízí GS také *citační vyhledávání*, známé ve vědeckém prostředí z profesionálních citačních databází Web of Science či Scopus. Výsledky (plné texty, abstrakty nebo jen bibliografické citace vědeckých prací) jsou sdruženy do skupin odpovídajících jednotlivým dokumentům a jsou setříděny podle relevantnosti a důležitosti vycházející z počtu citací a dalších kritérií.

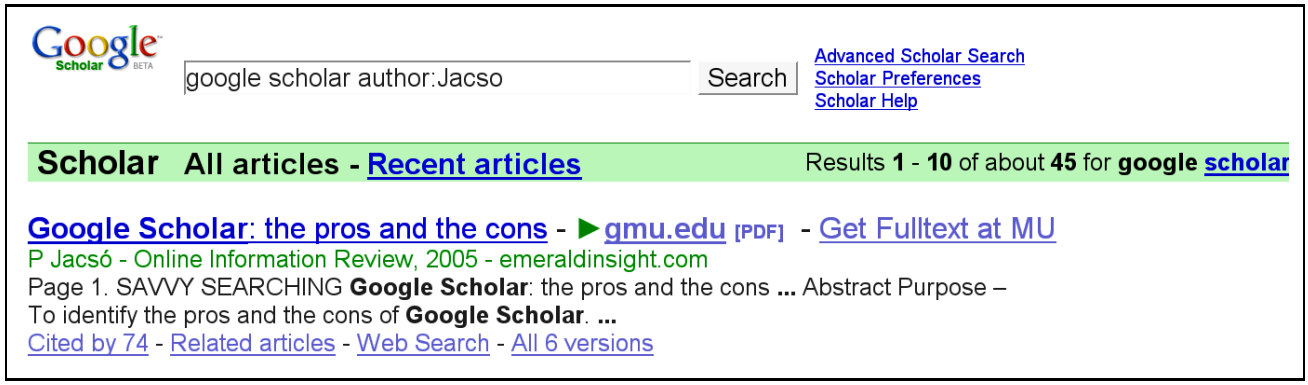
2 Začínáme s Google Scholar

Práce s GS začíná vyhledáváním – k dispozici je buď klasické jednoduché vyhledávací okénko pro zadání „klíčových" slov nebo vyhledávací formulář pro přesnější vyhledávání. Obojí je stejné resp. velmi podobné tomu, co známe z klasického vyhledávače Google. Ovšem výsledek vyhledávání již vypadá jinak. Stránka výsledků nezobrazuje přímo jednotlivé dokumenty, nýbrž seznam *prací*. Každá práce je tvořena jedním či několika různými *projevy* téhož díla. Například v případě článků může daný článek existovat (projevovat se) v podobě preprintu, článku publikovaného v recenzovaném časopise, dokumentu umístěného na osobním webu autora, souboru v univerzitním repozitáři, záznamu v abstraktové databázi nebo pouhé citace. Google Scholar se snaží (automatizovaně) seskupit všechny projevy daného článku dohromady, do jedné vědecké práce. K této práci se pak také vztahuje například počet citací – ten je dán jako součet citací všech projevů dané práce.

Každá práce má klikací název vedoucí na plný text či abstrakt, je-li dostupný on-line. Klikací název nemají samostatně stojící citace (Google Scholar je extrahuje z referencí, bibliografií, poznámek pod čarou apod.) a položky typu kniha.

Pro každou práci nabízí Google Scholar řadu funkcí:

- *Cited by* (Počet citací tohoto článku): uvádí práce, které citovaly danou práci a jsou obsaženy v databázi GS;
- *Related articles* (Související články): hledá další práce, které jsou tematicky podobné článkům dané práce;
- *Web Search* (Hledání na webu): vyhledává informace o dané práci na webu, přes klasický webový vyhledávač Google;
- *All x versions* (Všechny verze – počet: x): uvádí, z kolika projevů se skládá daná práce a umožňuje všechny projevy práce zobrazit.



Obrázek 1: Příklad jedné práce v seznamu výsledků Google Scholar

Kromě těchto základních funkcí mohou být nabídnuty i další funkce, například propojení prací na domovskou instituci uživatele, pokud toto propojení příslušná instituce v GS nastavila. Sem patří odkaz na projevy práce v katalogu domovské knihovny nebo odkaz na plný text práce v komerční plnotextové databázi, do níž má domovská instituce uživatele přístup (odkaz *Get Fulltext at MU* – viz dále v textu).

3 Přednosti a nedostatky Google Scholar

Google Scholar má své zastánce i odpůrce. Argumenty obou stran se pokusme shrnout v rozboru hlavních předností a nedostatků tohoto nástroje.

3.1 Přednosti

Mezi nejčastěji uváděné přednosti patří obrovský rozsah databáze a její univerzálnost, možnost citačního vyhledávání, které v některých případech poskytuje více výsledků než jiné tradiční nástroje, možnost zapojit do databáze i lokální vědecký obsah instituce, a v neposlední řadě také rychlé vyhledávání, snadnost použití a bezplatné využívání odkudkoliv.

Pokrytí, rozsah

Od svého vzniku prošel největším rozmachem samotný obsah databáze Google Scholar. Ta nyní představuje vůbec nejrozsáhlejší zdroj informací z vědeckého prostředí. Prakticky již každý významný světový producent vědecké literatury je zapojen nebo usiluje o zapojení svého obsahu do Google Scholar. Dokonce i takoví hráči jako Elsevier či Americká chemická společnost, kteří stáli zpočátku bokem, poskytují firmě Google obsah svých komerčních časopisů a databází k indexaci. Kromě zlepšeného časopiseckého pokrytí bylo výrazným impulsem pro rozvoj služby také zařazení dat z projektu Google Books. Umožnilo zpřístupnit pro vyhledávání plnotextové indexy miliónů digitalizovaných knih. Velkým kladem je také pokrytí značné části digitálních repozitářů a pre/post-printových archivů, zejména v oblastech přírodních věd, medicíny a počítačových věd (byť ne vždy již musí být zaindexován celý repozitář), což zlepšuje mimo jiné i přístup uživatelů k volně dostupné literatuře (open access). Vedle plnotextových zdrojů indexuje Google Scholar také milióny záznamů v bibliografických a abstraktových databázích od předních světových vydavatelů.

Univerzálnost

Google Scholar není specializován na nějakou konkrétní oblast či vědeckou disciplínu, i když jeho pokrytí jednotlivých oborů může být značně rozdílné (mezi nejlépe pokrytými se uvádí přírodní vědy, medicína a počítačové vědy). Široký záběr je velkou výhodou při vyhledávání témat, která se obvykle nevejdou do úzce specializovaných odborných databází, a při multidisciplinárním vyhledávání. Na rozdíl od mnoha jiných odborných zdrojů neomezuje také Google Scholar své geografické a jazykové pokrytí na anglosaský svět, ale indexuje vědeckou literaturu celosvětově a v různých jazycích.

Citační vyhledávání

Jak bylo již zmíněno v úvodní charakteristice, GS nabízí také funkci citačního vyhledávání („kdo cituje můj článek"). Umožňuje tím pohyb informačním prostorem v čase nejen vzad (přes seznam v článku použité literatury), ale i vpřed – od článku k novějším dokumentům, které z daného článku vychází a citují ho. Přestože stávající verze citačního vyhledávání v GS trpí některými neduhy (vyplývajícími hlavně z toho, že veškerá extrakce a analýza citací je prováděna výhradně jen strojovým způsobem, bez dodatečných ručních kontrol a korekcí chyb), pro řadu uživatelů představuje vítaný doplněk klasických komerčních citačních indexů, neboť díky obrovskému záběru databáze GS poskytuje často větší počet citací (byť ne vždy stoprocentně spolehlivých) než specializované citační databáze Web of Science a Scopus.

Napojení na domovskou instituci

V Google Scholar jsou dostupné nejen zdroje od velkých vydavatelů vědecké literatury. I běžné akademické knihovny mohou nabídnout obsah svých databází a katalogů k zařazení do databáze Google Scholar, takže uživateli mohou být při vyhledávání nabídnuty nejen zdroje na webu či u vydavatelů, ale i odkazy na relevantní literaturu dostupnou v jeho domovské knihovně (služba Library Search). S využitím technologie OpenURL mohou být také výsledky vyhledávání propojeny na plné texty článků v komerčních databázích zakoupených uživatelovou mateřskou institucí (služba Library Link). Tuto funkcionalitu jsme zprovoznili nedávno i pro Masarykovu univerzitu: uživatelé provádějící vyhledávání v GS z počítačů MU uvidí nyní vedle mnoha vyhledaných položek odkaz „Get Fulltext at MU", který je zavede přímo na plný text dokumentu v licencovaných fulltextových zdrojích dostupných pro MU.

Google navíc nabízí akademickým institucím digitalizaci jejich vlastních vědeckých časopisů (pokud například celý časopis nebo jeho starší ročníky neexistují v digitální podobě) a jejich začlení do databáze Google Scholar.

Snadnost použití, rychlost, cena

Podobně jako obecný vyhledávač Google nabízí i Google Scholar velmi snadný způsob zadávání vyhledávacích dotazů, který preferuje velká část uživatelů i v akademickém prostředí. Řada uživatelů nemá chuť učit se složité vyhledávací postupy a studovat sofistikovaná vyhledávací rozhraní. Ve spojení s okamžitou odezvou (Google Scholar využívá propracovaných technologií a zázemí mateřské firmy, které dokáží prohledávat bleskurychle jakkoliv rozsáhlou databázi) může uživatel získat často dostatečně kvalitní odpověď na to, co právě potřebuje. A navíc, vše je zadarmo – na rozdíl od stále rostoucích cen většiny komerčních informačních zdrojů a systémů.

3.2 Nedostatky

Hlavní výtky ke Google Scholar (a je nutno říci, že část informačních specialistů nemá Google Scholar právě v oblibě) spadají do tří oblastí – chyby softwaru, nedostatečně propracované pokročilé vyhledávání, a zejména nedostatek transparentnosti co se týče obsahu databáze a některých funkcí systému.

Chyby a nedostatky v softwaru

Jeden z nejznámějších kritiků GS, Peter Jacsó, charakterizuje v [1] problém následovně: „zatímco obecný vyhledávač Google odvádí obdivuhodnou práci ve světě nestrukturovaných webových stránek, software Google Scholar pokračuje v mizerných výsledcích při zpracování vysoce strukturovaných a metadaty opatřených vědeckých dokumentů". Už od uvedení Google Scholar (který vznikl vlastně jako vedlejší produkt resp. experiment v rámci sabbatical období některých vývojářů Google) byly systému vytýkány „školácké" chyby při analýze strukturovaných vědeckých dokumentů. Například parsovací algoritmy GS nedokáží vždy dostatečně spolehlivě identifikovat u jednotlivých dokumentů klíčové metadatové prvky – jako jména autorů (důsledkem jsou kuriózní autoři typu F. Password nebo V. Chapter), datum (někdy to vypadá, jakoby bylo za datum považováno téměř jakékoliv čtyřmístné číslo v dokumentu, což vede k paradoxům, kdy např. článek z počátku 20. století cituje články ze století 21.) nebo tematické zařazení příslušného dokumentu. To vede k nepřesnostem jednak při pokročilém cíleném vyhledávání

(pomocí jmen autorů, data, či s využitím tematických filtrů), zejména ale při citačním vyhledávání, neboť automatické párování citujících a citovaných dokumentů má k dokonalosti dost daleko. (Nutno ovšem přiznat, že vzhledem k chaosu a nezměrné variabilitě panující v oblasti citování, to nemají programátoři GS zrovna snadné). Nespolehlivé a někdy i značně nadsazené bývají kvůli tomu často i odhady počtu článků (počty hitů resp. počty citací).

Zásadní výtky kritiků nesměřují však ani tak k chybám samotným (i v oblasti relativně dobře strukturovaných vědeckých dokumentů je – s ohledem na již zmíněnou variabilitu a obrovský rozsah databáze – automatická extrakce metadat a citací úkol nesmírně obtížný), jako spíše k tomu, že většina chyb přetrvává v systému dlouhodobě. O dosavadní „intenzitě" vývoje nástroje Google Scholar leccos vypovídá i fakt, že čtyři roky po svém uvedení je stále ještě označován jako beta-verze.

Nedostatečné pokročilé vyhledávání

Kromě populárního jednoduchého vyhledávání známého z klasického vyhledávače Google nabízí Google Scholar i pokročilé vyhledávání. Uživatel má možnost například omezit vyhledávání na zadané autory (GS ovšem není schopen rozlišit mezi autory stejného jména), časový interval, zdroj (např. časopis) či tematickou kategorii (v české verzi GS není dostupné). Účinnost tohoto vyhledávání je však výrazně poznamenána problémy s ne vždy spolehlivou detekcí metadat (viz předchozí bod) případně s absencí důležitých metadatových prvků u části dokumentů. Navíc jsou možnosti pokročilého vyhledávání GS oproti specializovaným profesionálním databázím poměrně chudobné (což je z větší části daň za univerzální záběr GS). Chybí také možnost vlastního třídění výsledků.

Nedostatek transparentnosti

Zásadní nedostatek však spatřuje velká část kritiků v nedostatku informací o tom, co všechno vlastně Google Scholar indexuje (jaké je jeho pokrytí) a jaký je jeho rating algoritmus. Pro řadu uživatelů by bylo užitečné vědět, jak velká je databáze GS, které časopisy od kterých vydavatelů a za jaké období obsahuje, jaké je přesně pokrytí určité jazykové oblasti, které repozitáře a v jaké míře úplnosti jsou indexovány, jaká je frekvence aktualizace údajů apod. Bohužel, Google Scholar k tomu neposkytuje téměř žádné informace. Většina z toho, co bylo publikováno, vychází jen ze značně nespolehlivých, neúplných či zastaralých experimentů od uživatelů a recenzentů. Chybí i kritérium „vědeckosti", podle něhož jsou dokumenty do databáze zařazovány (v GS lze najít také editoriály, učebnice, studentské práce a další typy dokumentů, které nemají striktně vědeckou povahu).

Velmi neurčité je rovněž vyjádření vývojářů k tomu, jakým způsobem jsou řazeny výsledky vyhledávání. Je zřejmé, že řazení nemůže být založeno na page-rank algoritmu, jaký používá klasický vyhledávač Google. Zveřejněna byla jen vágní vyjádření v tom smyslu, že při stanovování relevance je do úvahy brána řada faktorů – včetně toho, kdo daný článek napsal, kde byl publikován, a především je (nějak) zohledňována citovanost článku. Tato velká míra neurčitosti mj. značně omezuje využití GS pro provádění hodnověrných citačních analýz a porovnání.

4 Shrnutí

Google Scholar je zajímavý a řadou uživatelů často využívaný nástroj pro vyhledávání vědeckých informací. Vzhledem k některým nedostatkům ve svém software však zatím nenaplňuje všechna očekávání vyvolaná při jeho uvedení před čtyřmi roky. Výhodou je obrovská databáze vědeckých informací a její využití zejména při multioborovém vyhledávání. Je neocenitelným pomocníkem pro nalezení rychlé odpovědi a při hledání „jehly v kupce sena". Pro hlubší a přesnější bádání dává však (zatím?) řada vědců přednost osvědčeným specializovaným oborovým databázím; to platí i pro citační analýzy.

Literatura

[1] Péter Jacsó. *Google Scholar revisited.* Online Information Review, Vol. 32, Iss. 1, 2008. Dostupné online též na `http://www.jacso.info/PDFs/`

jacso-GS-revisited-OIR-2008-32-1.pdf

[2] R. Vine. *Google Scholar. Electronic Resources Review.* J Med Lib Assoc, January 2006. http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pmcentrez&artid=1324783

[3] Tamar Sadeh. *Google Scholar versus Metasearch Systems.* High Energy Physics Libraries Webzine, issue 12, February 2006. http://library.cern.ch/HEPLW/12/papers/1/

[4] *An interview with Anurag Acharya, Google Scholar lead engineer.* Google Librarian Central – Article 12/2006. http://www.google.com/librariancenter/articles/0612_01.html □