Získávání dat z bibliografických databází

Dalibor Fiala

Katedra informatiky a výpočetní techniky
Západočeská univerzita v Plzni
Univerzitní 8, 306 14 Plzeň

dalfia@kiv.zcu.cz

Abstrakt. Známé bibliografické databáze splňují několik funkcí a mohou v zásadě sloužit jako vyhledávače odborných publikací, citační indexy nebo kalkulačky bibliometrických indikátorů. Mezi nejznámější z nich patří Web of Science, Scopus, ACM Digital Library, DBLP, CiteSeer^X a Google Scholar. Větší množství dat z těchto databází je možno s úspěchem použít mj. pro bibliometrická měření, analýzu citačních sítí a sítí spolupráce a pro vizualizaci produktivity a kvality vědeckého výzkumu. Tato data lze v některých případech získat pouze ručně, ale v jiných i automatizovaně. V tomto příspěvku si uvedeme přehled bibliografických databází a možnosti získávání dat z nich a podrobněji se zaměříme na databáze Web of Science a Scopus.

Klíčová slova: WoS, Scopus, DBLP, CiteSeer, ACM DL, Google Scholar.

1 Úvod

V tomto příspěvku budeme kvůli zjednodušení nazývat bibliografickou databází každý systém umožňující v různé míře vyhledávání odborných publikací, poskytování bibliografických informací o nich, procházení referencí a citací a přístup k jejich abstraktům nebo dokonce plným textům. V zásadě je tedy podle funkcionality můžeme rozdělit do tří vzájemně se nevylučujících skupin: vyhledávače odborných publikací, citační indexy a „kalkulátory" bibliometrických indikátorů. Databáze v první skupině umožňují typicky vyhledávání mj. v názvech článků, jménech autorů i jejich adresách, klíčových slovech, abstraktech a někdy i plných textech. Citační indexy poskytují především možnost nalézat citované a citující publikace a sledovat tak vývoj nějaké úzce vymezené problematiky určitého vědního oboru. Poslední kategorie bibliografických databází je zaměřena na měření nejrůznějších aspektů publikační činnosti, např. počty publikací a citací, h-index, impaktní faktor aj.

Mezi nejznámější bibliografické databáze patří Web of Science (dříve ISI, Thomson Reuters, nyní Clarivate Analytics [1]), Scopus [2], ACM Digital Library (rozšířená verze známá jako ACM Portal nebo Guide [3]), DBLP [4], CiteSeer^X (dříve CiteSeer [5]) a Google Scholar [6]. Některé jejich vlastnosti přehledně shrnuje tabulka 1 a jimi se budeme dále zabývat. Mimo ně samozřejmě existují i mnohé další, které

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 72-76.

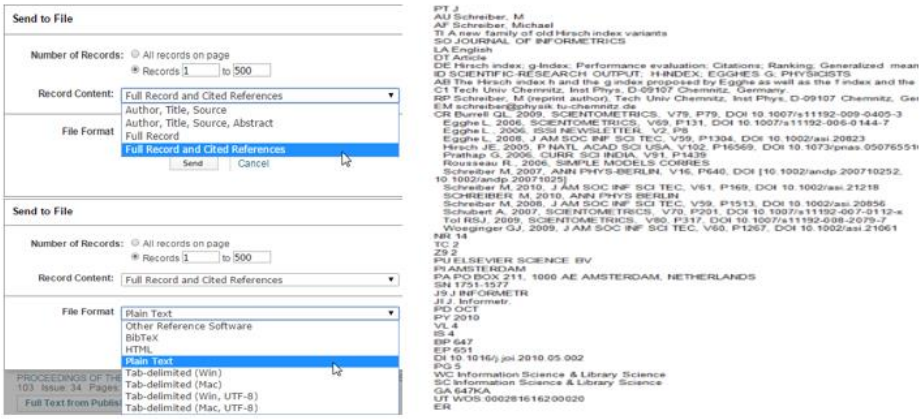
nejsou tématem tohoto příspěvku, jako např. IEEE Xplore [7], PubMed [8], arXiv [9], Microsoft Academic [10], SemanticScholar [11] atd.

Tab. 1. Tabulka vlastností vybraných bibliografických databází (červen 2017)

	ACM DL (Guide)	CiteSeer^{X}	DBLP	Google Scholar	Scopus	Web of Science
Zdarma	částečně	ano	ano	ano	ne	ne
Automatizovaný	ne	ano	ne	ano	ne	ne
Počet záznamů	2,68 mil.	10 mil.	3,81 mil.	100+ mil.	75 mil.	62,25 mil.
Vše ke stažení	ne	ano	ano	ne	ne	ne
Propojení referencí	ano	ano	ne	ne	ano	ano
Propojení citací	ano	ano	ne	ano	ano	ano
Počet citací článku	ano	ano	ne	ano	ano	ano
Počet citací autora	ano	nepřímo	ne	nepřímo	ano	ano
Vědní obor	informatika	informatika	informatika	všechny	všechny	všechny

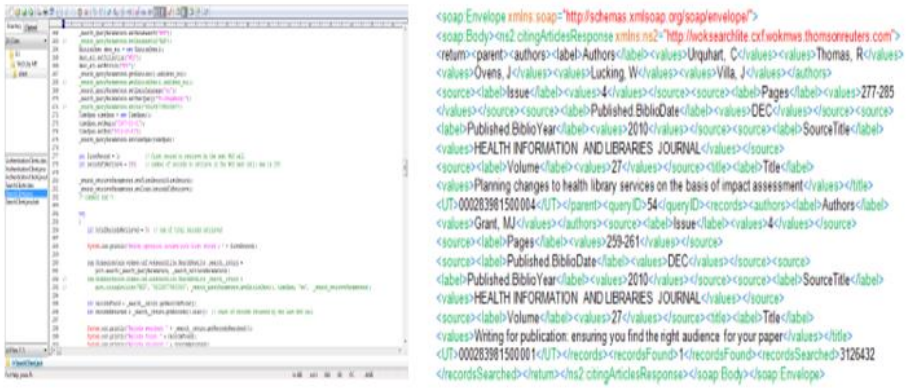
2 Export dat

Některé uvedené databáze jsou zdarma a poskytují dokonce všechna svá data ke stažení ve formě jednoho obřího souboru XML (DBLP) nebo na vyžádání jako celý *dump* databáze MySQL (CiteSeer^{X}). Google Scholar, jenž je rovněž zdarma, žádnou takovou možnost nenabízí a dokonce (zřejmě záměrně) neposkytuje ani žádné programátorské rozhraní (API) pro přístup ke svým datům. To samé platí i pro částečně zpoplatněnou ACM DL. V obou případech by tedy získání většího množství dat bylo možné pouze programově automatizovaným webovým pavoukem (robotem) se všemi problémy a omezeními s tím spojenými. U dvou zbývajících placených databází Web of Science (WoS) a Scopus je situace pochopitelně odlišná. Na obr. 1 je vidět ruční export bibliografických záznamů (maximálně 500 najednou) do čistého textu ve Web of Science:



Obr. 1. Ruční export záznamů z databáze Web of Science

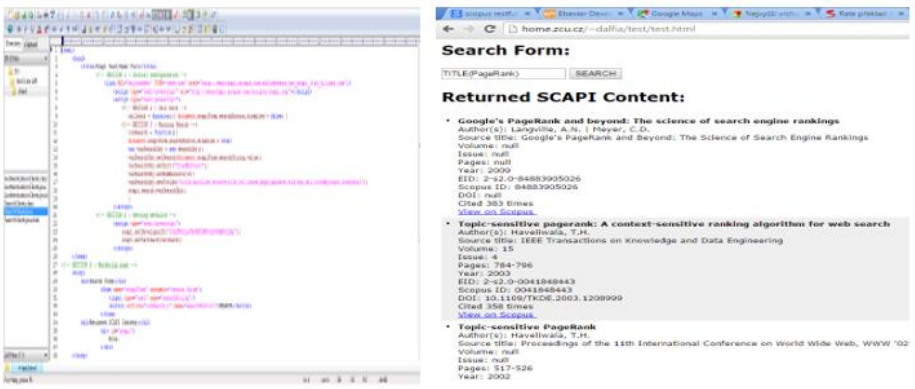
Omezení počtu najednou exportovaných záznamů je i ve Scopusu: 2000. Pokud se spokojíme s jen minimalistickými informacemi o publikacích (v zásadě jen název, autor, místo a rok vydání), zvyšuje se tento limit až na 20000 ve Scopusu a nově též na 5000 ve WoSu. Pro export řádově vyšších počtů bibliografických záznamů bude tedy vhodné využít API, které obě databáze nabízejí, a záznamy stahovat automaticky pro tento účel vytvořeným programem. Na obr. 2 je ukázka programového kódu a výsledného kusu stažených dat ve formátu XML, jak ho poskytovala starší „odlehčená“ verze WoS API, tzv. *Web Services Lite*.



Obr. 2. Automatický import záznamů z databáze Web of Science (*Web Services Lite*)

Kromě této odlehčené verze je rovněž k dispozici API *Web Services Expanded* poskytující i mnoho dalších údajů o článcích jako jsou např. počty citací, adresy autorů nebo abstrakty [12]. Všech rozhraní lze užívat jen po registraci a získání přístupového klíče. Podobně na obr. 3 je ukázka programového kódu, který přes Scopus API [13],

kdysi označované jako *SCAPI*, zobrazuje importované záznamy ve webové aplikaci, jež je na rozdíl od WoS API povinným cílovým místem využití stahovaných dat:



Obr. 3. Automatický import záznamů z databáze Scopus (*SCAPI*)

3 Závěr

Dosud nezmiňovanou možností získání dat je jejich nákup přímo od provozovatele databáze. Tuto možnost sám autor tohoto příspěvku vyzkoušel u databáze Web of Science. Výhodou je dodání dat „na míru“ podle přesně zadaných kritérií ve formátu XML. Nevýhodou je vysoká cena, a to i pro pracovníka instituce, která je standardním předplatitelem této databáze a má k ní přístup přes webové rozhraní. V každém případě mají získaná data cenné využití, ať už v bibliometrických studiích, grafových analýzách nebo obecně jakýchkoliv jiných experimentech nad rozsáhlými daty.

Literatura

1. Web of Science: http://clarivate.com/scientific-and-academic-research/research-discovery/web-of-science/. Získáno 28. 6. 2017.
2. Scopus: https://www.scopus.com/. Získáno 28. 6. 2017.
3. ACM Digital Library: http://dl.acm.org/. Získáno 28. 6. 2017.
4. DBLP: http://dblp.org/. Získáno 28. 6. 2017.
5. CiteSeer^{X}: http://citeseerx.ist.psu.edu/. Získáno 28. 6. 2017.
6. Google Scholar: https://scholar.google.com/. Získáno 28. 6. 2017.
7. IEEE Xplore: http://ieeexplore.ieee.org/. Získáno 28. 6. 2017.
8. PubMed: https://www.ncbi.nlm.nih.gov/pubmed/. Získáno 28. 6. 2017.
9. arXiv: https://arxiv.org/. Získáno 28. 6. 2017.
10. Microsoft Academic: https://academic.microsoft.com/. Získáno 28. 6. 2017.
11. Semantic Scholar: https://www.semanticscholar.org/. Získáno 28. 6. 2017.
12. Web of Science Data Integration: http://ip-science.interest.thomsonreuters.com/data-integration/. Získáno 28. 6. 2017.
13. Scopus APIs: https://www.elsevier.com/solutions/scopus/features/api/. Získáno 28.6. 2017.

Poděkování: Tato publikace byla podpořena projektem LO1506 Ministerstva školství, mládeže a tělovýchovy ČR.

Annotation:

Data Acquisition from Bibliographic Databases

The established bibliographic databases have a number of functionalities and can, in principle, serve as search engines of academic papers, citation indices, or calculators of bibliometric indicators. Web of Science, Scopus, ACM Digital Library, DBLP, CiteSeer^{X}, and Google Scholar belong to the best known ones. Larger amounts of data from these databases can be successfully used for bibliometric measurements, citation and collaboration networks analysis, and for the visualization of the production and quality of scientific research. These data can be acquired only manually in some cases but also automatically in some others. In this short paper we will give an overview of bibliographic databases and the possibilities of data acquisition from them and we will focus in more detail on Web of Science and Scopus.