Zkušenosti s využitím EIZ pro hodnocení vědy

Pavel Mika
Knihovna Akademie věd ČR
mika@knav.cz

INFORUM 2015: 21. ročník konference o profesionálních informačních zdrojích
Praha, 26. - 27. 5. 2015

Abstrakt

Příspěvek představí základní možnosti využití elektronických informačních zdrojů pro hodnocení vědy z hlediska bibliometrie, tedy kvantitativních metod analýzy dat.

Informační systémy CRIS a komerční citační databáze obsahují informace využitelné pro hodnocení jednotlivých vědců, vědeckých pracovišť a organizací, ale i celých vědních oborů nebo regionů v rámci světového porovnání. Myšleny jsou informace dostupné na základě záznamů o vědeckých výstupech (především článcích, ale i dalších) obsažených v těchto informačních zdrojích.

Informace získané analýzou uložených dat – bibliografických záznamů, případně citačních vazeb mezi nimi mohou vypovídat o postavení hodnocené jednotky z hlediska počtu, vědeckého ohlasu a částečně i kvality (měřené citovaností) jednotlivých druhů vědeckých výsledků a jejich podíl na domácí, světové produkci. Nejzajímavější informace mohou poskytnout právě data o citačních vazbách. Dále je možné sledovat oborové zaměření či přímo aktuální témata pomocí analýzy klíčových slov. Analýzou autorské spolupráce získáme pohled na mezinárodní propojení vědního prostoru.

Jednotlivé informační zdroje nabízejí různé možnosti stažení dat pro účely jejich analýzy, popřípadě umožňují přímo analýzu dat a jejich následné zobrazení v přehledové formě pomocí tabulek a grafů. Práce se staženými daty je z jedné strany náročnější, ale umožňuje například lepší kontrolu dat než přímá analýza v informačním systému, jak bude představeno v příspěvku.

Úvod

Zkušenosti popisované v tomto článku jsou nasbírány při práci na několika interních analýzách v rámci Akademie věd České republiky, samotném zkoumání využitelnosti různých zdrojů dat, pomoci při přípravě bibliometrických informací v rámci Hodnocení výzkumné a odborné činnosti pracovišť AV ČR za léta 2010-2014[1] a v rámci zapojení do projektu Efektivní systém hodnocení a financování výzkumu, vývoje a inovací (IPN Metodika)[2], především v právě probíhajícím pilotním ověření nově vznikající metodiky.

Bibliometrie poskytuje kvantitativní informace o vědeckých výstupech. Tyto informace mohou odrážet kvalitu těchto vědeckých výstupů a to především pomocí citační analýzy. Bibliometrické informace by ovšem měly sloužit především jako podkladové informace a neměly by být využívány k přímému rozhodování například o financování dané výzkumné instituce. Ideální kombinací pro

[1] http://interni.avcr.cz/Informace_pro_pracoviste/informace-o-hodnoceni-pracovist-av-cr/hodnoceni-vyzkumne-a-odborne-cinnosti-pracovist-av-cr-za-obdobi-2010-2014/index.html
[2] http://metodika.reformy-msmt.cz/

hodnocení vědy se jeví využití peer review metody, tedy hodnocení odborníky z daného oboru, doplněné o bibliometrické informace.

Předpokladem pro tvorbu bibliometrických analýz jsou data o vědeckých výstupech získaná z některého z dostupných informačních zdrojů. Těmito zdroji jsou obvykle citační bibliografické databáze nebo vědecké informační systémy – tzv. CRIS (Current Research Information System).

Bibliometrické indikátory

V následující části článku je uveden výčet několika bibliometrických indikátorů, které je možné získat z dat z citačních databází nebo CRIS systémů. Vzhledem k tomu, že se tento článek nezabývá přímo bibliometrickými indikátory, není jejich popis podrobný.

Značnou výpovědní hodnotu mají indikátory založené na počtu citací. Základem citačních analýz je předpoklad, že citace poukazuje na jinou práci především z pozitivního hlediska, tedy jako na zdroj informací důležitých pro současnou práci a tím mu tedy přidává na důležitosti. Může však nastat případ, kdy je citace uvedena v rámci polemiky či přímo nesouhlasu s odkazovanou prací. Citace mohou mít také různou váhu – na jedné straně případ, kdy je citace uvedena mezi několika dalšími jako základní zdroj informací k dané tématice, v druhém případě je citována jediná práce kvůli jejím současným objevům. Tyto rozdíly bohužel není možné kvantitativní analýzou nijak postihnout, a proto se pracuje s výše napsaným předpokladem, že citace odpovídá pozitivnímu ohlasu.

Citační zvyklosti se liší v různých vědeckých oborech a nelze tedy srovnávat samotné počty citací mezi pracemi například z matematiky (obecně málo citovaný obor) oproti pracím se zaměřením na chemii (více citovaný obor). Pro účely takového porovnání napříč obory dochází k normalizaci počtu citací oproti průměrnému počtu citací.

Základními dvěma citačními indikátory jsou oborově normalizovaný citační impakt (Mean Field Normalized Citaion Impact – MFNCI, v praxi jsou používány i jiné názvy) a podíl na nejcitovanějších (obvykle 10% či 25%) pracích v oboru (proportion of top 10% publications – PP10%).

MNFCI je vypočítán jako podíl počtu citací publikace a průměrného počtu citací všech publikací ze stejného oboru, stejné roku vydání a stejného typu dokumentu (článek, review). V případě více oborů článku se vypočte podíl za každý obor a bere se jejich průměr. Pro určení hodnoty tohoto indikátoru u skupiny více publikací, obvykle představujících vědeckou produkci sledované jednotky, se bere průměrná hodnota za jednotlivé publikace. Tento indikátor je jednoduše interpretovatelný, jelikož hodnota 1 představuje světový průměr (za předpokladu, že svět je vše co je obsaženo v databázi) a např. hodnota 1,45 znamená, že daná skupina prací překračuje průměrnou citovanost o 45 procent. Oproti této výhodě je indikátor snadno ovlivnitelný jediným vysoce citovaným článkem a tak může být zkreslena jeho výpovědní hodnota. Vzhledem k tomu, že indikátor pracuje právě s průměrem, měl by být založen na dostatečném množství poměřovaných publikací.

PP10% (25%) vyjadřuje podíl prací, které se řadí mezi nejcitovanější práce v oboru za daný rok a typ dokumentu. Tento indikátor je vhodný ke sledování kvality a nedochází u něj ke zkreslení při zařazení jednoho vysoce citovaného článku do souboru méně citovaných.

Indikátor sledující poměr podílu výstupů vzniklých v mezinárodní či národní spolupráci, může vypovídat o zapojení do širšího vědeckého prostoru a u menších zemí, jako je Česká republika, často koreluje s citačními indikátory – čím vyšší podíl výstupů se zahraniční spoluprací, tím je větší citovanost daného souboru dokumentů.

Citovanost zdrojového časopisu je možné vyjádřit několika indikátory. Nejznámější je tzv. Impakt faktor (Journal Impact Factor), který však není normalizovaný. Sofistikovanějšími indikátory sledující citovanost časopisů jsou například SNIP či AIS.

Citační databáze

Na poli citačních databází jsou hlavními hráči dva rovnocenní konkurenti: databáze Web of Science produkovaná firmou Thomson Reuters a databáze Scopus od firmy Elsevier.

Databáze Web of Science (WoS), součást systému Web of Knowledge, pokrývá produkci více než 12 000 mezinárodních i místních odborných časopisů v každé z oblastí přírodních, sociálních a humanitních věd a umění. Její pokrytí časopisů je rozsáhlé, nesnaží se však o úplnost. Naopak časopisy jsou do databáze vybírány na základě několika kritérií[3], kterými jsou: publikační standardy (peer review, pravidelné vycházení), obsah (přidaná hodnota v daném oboru), rozmanitost (mezinárodní, regionální vliv autorů, redaktorů), citační analýza (předcházející vědecké práce redaktorů a autorů) a jazyk (většina fulltextů v angličtině). Publikování článku v časopisu, který je indexován v databázi WoS, tedy může představovat jistou míru kvality.

Při využití databáze WoS pro bibliometrické analýzy je nutné brát v úvahu právě její výběrovost a tedy to, že obvykle není zahrnuta veškerá vědecká produkce sledované jednotky.

Databáze Scopus pokrývající více než 21 000 titulů periodik od 5000 vydavatelů z celého světa se naopak snaží o širší pokrytí celosvětové vědecké produkce.

Porovnání pokrytí časopisů v rámci ČR těmito dvěma databázemi bude podrobeno dalším analýzám, jelikož právě výběr zdroje dat je jedním z klíčových rozhodnutí v rámci kvantitativního hodnocení vědy.

Obě dvě citační databáze nabízejí nadstavbové analytické nástroje, InCites v případě WoS a SciVal pro databázi Scopus. Tyto nástroje umožňují jednoduchou formou vytvoření bibliometrických analýz sledujících různé indikátory.

V rámci dosavadních pracovních příležitostí jsem měl možnost pracovat především s daty a nástroji firmy Thomson Reuters. Proto bych zde rád krátce představil možnosti jejich využití. Zpracování dat a informací z databáze Web of Science, případně Journal Citation Reports (JCR) nebo z nástroje InCites bych rozdělil právě podle zdroje, respektive způsobu získání dat na tři možnosti: práce přímo s databází WoS/JCR případně s ručně staženými daty, zpracování „surových“ dat a využití analytického nástroje InCites.

Pomocí uživatelského přístupu do databáze WoS je možné získat velké množství dat pro výpočet výše zmíněných indikátorů. Hlavním problémem je značná pracnost a časová náročnost získání a zpracování dat. Analýza oborů, druhů dokumentů či spolupracujících zemí je možná přímo pomocí nástroje Analyze results. Základní citační analýza je možná u počtu dokumentů do 10 000, přičemž získáme pouze informaci o počtu citací a autocitací. Není zde použita žádná normalizace a pokud bychom chtěli získat podkladová data pro výpočet MFNCI, tedy průměrný počet citací všech publikací v oboru, jsme omezeni právě počtem 10000. Pro PP10% je možné získat podkladová data pro jeden obor pomocí několika dotazů. Získání dat pro větší množství oborů už však může představovat určité

[3] TESTA, Jim. The Thomson Reuters journal selection process. [online]. 2012 [cit. 2014-10-30]. Dostupné z: http://wokinfo.com/essays/journal-selection-process/

obtíže, především vzhledem k časové náročnosti. Také samotné stažení dat o publikacích je omezeno maximálním počtem záznamů pro jedno stažení, který je 500, čímž se proces stažení většího objemu dat komplikuje.

Pokud máme k dispozici surová data, tedy podkladová data včetně informací o počtu citací nemusí být zpracování bibliometrické analýzy tolik náročné na ruční práci. Je zde však omezení z hlediska rozsahu a stáří dat. Surová data bývají dodávána za určité časové období, dále omezena jen například na jednu zemi a jsou výpovědná pouze k jejich datu dodání na rozdíl od „živé" databáze. Regionální omezení dat, tedy přístup pouze k datům o publikacích obsahujících Czech republic v adrese, znemožňuje získání jakýchkoliv podkladových dat pro výpočet normalizovaných citačních indikátorů. Tyto světové průměry citací v oboru a minimální hodnoty citací pro PP10%, souhrnně *benchmarks,* musejí být tedy dodány společně s daty. Výpočet indikátorů ze surových dat umožňuje více prostoru pro vyčištění dat, ověření jejich kvality a využití různých metod výpočtu indikátoru. Samozřejmě zpracování dat je také časově náročné.

Oproti tomu využití analytického nástroje InCites je časově nenáročné, stačí pouze několikrát kliknout a získáme přehledné tabulky a grafy. Nevýhodou může být možnost využití pouze přednastavených indikátorů, kde například u PP10% není možné jistit proporci v rámci PP25%. Hlavním problémem je ovšem čistota a definování podkladových dat. Tedy určení, které publikace patří ke sledované jednotce, samotné definování jednotek – univerzit, fakult, výzkumných organizací či přímo vědeckých týmů. Takovéto definování je nemožné pouze na straně uživatele a je zde nutná spolupráce s dodavatelem databáze. Pokud bychom chtěli využít pouze organizace definované v databázi WoS pomocí rejstříku Organisation-enhanced, jsme omezeni jednak malou granuitou (rozdělení pouze na úrovni celých univerzit či pouze AV ČR a nikoliv jejích ústavů) a také ne vždy správným přiřazením publikací.

Otázka definování souboru sledovaných publikací je samozřejmě na místě i v případě práce se surovými daty. Zde mohou velice dobře posloužit data z lokálních CRIS systémů, kde můžeme snáze identifikovat hodnocené jednotky a jejich publikace a poté je propojit s daty z databáze WoS, abychom získali údaje o citovanosti. Toto propojení dat na úrovni jednotlivých publikací je možné jednak, pokud jsou v CRIS systému ručně doplněny identifikátory WoS záznamů nebo pomocí porovnání záznamů na základě shodnosti, či podobnosti (kvůli překlepům, interpunkci apod.) názvu, zdrojového časopisu, roku vydání a jmen autorů.

CRIS systémy

Na národní úrovni slouží jako CRIS Informační systém výzkumu, experimentálního vývoje a inovací (IS VaVaI) Jeho úkolem, definovaným zákonem č. 130/2002 Sb., o podpoře výzkumu, experimentálního vývoje a inovací z veřejných prostředků a o změně některých souvisejících zákonů (zákon o podpoře výzkumu, experimentálního vývoje a inovací), ve znění pozdějších předpisů (dále jen zákon č. 130/2002 Sb.), je shromažďovat, zpracovávat, poskytovat a využívat údaje o výzkumu, experimentálním vývoji a inovacích (VaVaI) podporovaných z veřejných prostředků.

IS VaVaI v současné době tvoří celkem pět navzájem provázaných a veřejně dostupných částí, kterými jsou Centrální evidence aktivit výzkumu, experimentálního vývoje a inovací, Centrální evidence projektů, Rejstřík informací o výsledcích a Evidence veřejných soutěží ve výzkumu, experimentálním vývoji a inovacích. S ohledem na kontinuitu údajů o poskytované veřejné podpoře na VaVaI, je součástí IS VaVaI též Centrální evidence výzkumných záměrů. Právě Rejstřík informací o výsledcích (RIV), často zaměňovaný s celým IS VaVaI, obsahuje nejvyužitelnější data ke kvantitativním analýzám.

Data z IS VaVaI, která jsou veřejně k dispozici a jednoduše stažitelná umožňují základní analýzy sledující především publikační profil zkoumané jednotky. Je možné určit zapojení do výzkumného prostoru v rámci ČR pomocí podílu výstupů jednotky oproti všem výstupům a to nejlépe v rozdělení na jednotlivé vědní obory (IS VaVaI pracuje s vlastní, již trochu zastarávající, klasifikací složenou ze 123 oborů). Jak bylo popsáno výše velice dobré využití dat je pro definování výzkumných jednotek a následné propojení s WoS daty pro získání citační analýzy. Uvedení identifikátoru WoS záznamu v datech z IS VaVaI není povinné, ovšem toto pole v záznamu existuje a je tedy možné jej vyplnit a přibližně u 50% záznamů za několik posledních let je vyplněno.

Akademie věd ČR má svůj vlastní cenný informační zdroj, databázi ASEP (Automatizovaný systém evidence publikací) spravovaný Knihovnou AV ČR z dat dodávaných jednotlivými ústavy AV ČR. Od počátečního sběru bibliografických záznamů o publikačních výsledcích ukládaných v databázovém systému CDS/ISIS (bezplatně poskytovaného UNESCO) došlo postupně k přechodu na profesionální systém ARL (Advanced Rapid Library). ASEP dnes obsahuje více než čtvrt miliónu záznamů o publikačních i nepublikačních výstupech vědeckého výzkumu v AV ČR a je využíván ve dvou rovinách: jako katalog, v současnosti již nazývaný Repozitář AV ČR, unikátní informační systém pro ústavy AV ČR i pro jejich vedení, s možností ukládání plných textů a sloužící jako zdroj pro vytvoření výstupních sestav pro RIV – součást IS VaVaI, dále pak jako Analytika ASEP poskytující bibliografické přehledy a grafické výstupy.

Oproti analýze dat z IS VaVaI umožňuje ASEP například zjištění podílu prací v zahraniční spolupráci, jemnější rozdělení druhu dokumentů (rozdělení na články v impaktovaných časopisech) či definované samotné výzkumné týmy v rozhraní Analytik ASEP.

Etické aspekty na závěr

Kvantitativní analýza vědeckých výstupů může poskytnout cenné informace, které je možno využít pro potřeby hodnocení vědy. Tyto informace je třeba však správně interpretovat a zamezit zjednodušování závěrů z nich vycházejících. V roce 2012 byly sepsány doporučení k otázce hodnocení vědy pod názvem San Francisco Declaration on Research Assessment[4]. Jedno z hlavních doporučení bylo nepoužívat Impakt faktor k hodnocení kvality vědeckých článků, kromě toho pak bylo například doporučeno být transparentní a poskytovat data, na jejichž základě byly analýzy vytvořeny.

Potřeba dalších biblioemtrických „standardů“, která byla probírána na panelové diskusi v rámci konference STI2014 v Leidenu, vyústila ve formulování 10 bodů pod názvem Leiden Manifesto[5]. Prvním z těchto bodů je, že kvantitativní hodnocení by mělo podporovat kvalitativní a ne ho nahrazovat. Dalším důležitým bodem je poskytnutí možnosti ověření dat ze strany hodnocených ke zvýšení kvality těchto dat. Ostatní body jsou k nalezení v odkazovaném článku. Je důležité podotknout, že tato doporučení, často mířená přímo na producenty bibliometrických analýz je potřeba uvést do praxe a seznámit s nimi i širší okolí, kterých se týká, tedy především vědní management.

Kvantitativní hodnocení vědy může poskytnout cenné podkladové informace. Výběr zdroje dat a poté především jejich ověření a čištění jsou kritickou částí procesu zpracování bibliometrických analýz. Využití analytických nástrojů od producentů citačních databází usnadňuje získání bibliometrických indikátorů ovšem klade velké nároky na správné namapování a segmentaci dat. K tomuto účelu, nejen

[4] http://www.ascb.org/dora/

[5] HICKS, Diana, Paul WOUTERS, Ludo WALTMAN, Sarah de RIJCKE a Ismael RAFOLS. Bibliometrics: The Leiden Manifesto for research metrics. *Nature*. 2015, roč. 520, č. 7548, s. 429-431 [cit. 2015-04-28]. DOI: doi:10.1038/520429a. Dostupné z: http://www.nature.com/news/bibliometrics-the-leiden-manifesto-for-research-metrics-1.17351

však při práci s analytickými nástroji ale i přímo s daty, mohou dobře posloužit vědecké informační systémy.