BIG DATA SKRYTÁ V KNIHOVNÁCH V KONTEXTU MEDICÍNSKO-INFORMAČNÍM

Anna Motejlková, Richard Papík

Anotace

Snad ve všech odvětvích lidského působení se v současné době uplatňují tzv. big data vyznačující se extrémně velkým objemem dat, extrémně vysokou rychlostí, s jakou data vznikají, a extrémně velkou rozmanitostí dat. Velké soubory dat patřící do této kategorie se skrývají i v knihovnách. Kde přesně lze v knihovnách big data nalézt a jakým způsobem by jejich analýzy mohly být přínosné pro rozvoj vědy a výzkumu? Na tuto otázku se snaží odpovědět následující příspěvek.

Klíčová slova

big data, informační zdroje, rešeršní strategie

1. Úvod

V současné době rychle rostoucího množství dostupných informací je stále častěji třeba pracovat s velkými soubory dat. Této oblasti byl přisouzen termín „big data" (dále v textu bude užíván v původní anglické podobě, ve které se dostává i do českého odborného textu). Tento pojem označuje soubor dat, jehož velikost znemožňuje jeho uchovávání či zpracování za pomoci běžně dostupného hardwaru a softwaru. Vedle samotného rozsahu jsou tato data charakterizována rychlostí jejich vzniku, big data netvoří soubor statických dat ale naopak velmi rychle se vyvíjejících.

Big data vznikají dnes v téměř každém odvětví. Ačkoli to nemusí být na první pohled patrné, také knihovny disponují značným množstvím dat spadajících do této kategorie. Většina informací, které můžeme v knihovnách obvykle nalézt, bývá ukládána pomocí záznamů v některém z dostupných metadatových formátů. Tato alespoň částečná uspořádanost jim dává velkou výhodu, neboť je dělá snáze zpracovatelnými. Otázkou zůstává, zda mohou být velká data dostupná v knihovnách užitečná také v oblasti vědy a výzkumu.

Při snaze zmapovat současný stav problematiky big data v knihovnách narážíme na to, že v dostupných informačních pramenech se zatím příliš konkrétních zmínek na toto téma nevyskytuje. Není však pochyb o tom, že i v knihovnách, stejně jako ve většině moderních organizací, nalezneme velké množství různých dat spadajících do big data, které by stálo za to zpracovávat a vytěžovat z nich informace. Která data to jsou a jaké užitečné informace bychom z nich mohli získat, na to se zaměřuje následující příspěvek. Cílem tohoto příspěvku není výčet všech dostupných big data v knihovnách, ale nalezení těch dat, která by mohla potenciálně pomoci ve vědě a výzkumu.

2. Big data

Jako big data bývají označovány soubory dat vyznačující se extrémně velkým objemem dat, extrémně vysokou rychlostí, s jakou data vznikají, a extrémně velkou rozmanitostí dat. Velikost těchto souborů obvykle znemožňuje jejich uchovávání či zpracování za pomoci běžně dostupného hardwaru a softwaru. Hlavním úkolem big data je umožnit shromažďování, ukládání, správu a manipulaci s obrovským množstvím různorodých dat, a to správnou rychlostí a ve správný čas pro získání těch nejlepších poznatků, analýz či reakcí v reálném čase. [1]

Častým příkladem využití big data je zachycování a sledování údajů v reálném čase, což bývá typicky spojeno i s okamžitou vizualizací těchto dat. Na základě analýz množství aktuálních dat je pak možné predikovat budoucí chování dané skupiny, organizace tak mohou učinit lepší rozhodnutí a vytvářet plány do budoucnosti. Díky analýzám rozsáhlých souborů dat můžeme z dostupných informací vytěžovat další, které nemusejí být na první pohled patrné.

3. Role akademických / vědeckých knihoven

Abychom dokázali v knihovnách nalézt potenciálně užitečná data pro rozvoj vědy a výzkumu, je třeba si nejprve uvědomit, jaké postavení zaujímají knihovny v akademickém a vědeckém prostředí. Jedním z hlavních cílů a úkolů knihovny je budovat, udržovat a poskytovat přístup ke kvalitním a důvěryhodným zdrojům informací. Knihovny bývají vnímány jako jakési brány k odborné literatuře. Zároveň se od knihovny očekává spravování rozpočtu přiděleného na nákup informačních zdrojů, jak v tištěné tak v elektronické podobě. [2] Při hledání big data v knihovnách se proto zaměříme především na oblast informačních zdrojů.

4. Big data v knihovnách

Jak již bylo zmíněno, jedním z hlavních úkolů knihoven je spravovat fond informačních zdrojů. Podívejme se tedy nyní, kde v této oblasti můžeme nalézt potenciálně využitelné big data.

4.1 Publikační činnost

Výraznou oblastí v knihovnách s množstvím dostupných velkých souborů dat je publikační činnost. Knihovny si musí udržovat přehled o současné literární produkci, jaké knihy vychází, které nové časopisecké tituly jsou vytvářeny a které naopak zanikají, a jaké vznikají nové online nástroje, to vše na lokální i celosvětové úrovni. Jsou sledovány ediční plány jednotlivých vydavatelů, nabídky dodavatelských společností a také nejnovější trendy v oblasti techniky pro zpřístupnění informačních zdrojů čtenářům knihovny v pro ně co nejpřívětivější podobě. Dalším zdrojem informací tištěné i elektronické produkce jsou dokumenty, které získávají některé knihovny, včetně Národní knihovny České republiky, formou tzv. povinného výtisku. Jeho role je velmi významná nejen z pohledu konzervační funkce, ale i budoucího kompletního a komplexního

pohledu na dokumentovou komunikaci společnosti z perspektivy historického a kulturního dědictví. Zachycení publikační činnosti ve vědě, výzkumu a inovacích se děje rovněž prostřednictvím přirozené schopnosti knihoven vytvářet digitální archivy a digitální knihovny. [3]

4.2 Digitalizace

V posledních letech se velmi rozmáhá proces digitalizace, kdy se především starší díla, k nimž není dostupná elektronická verze, převádějí do digitální podoby pro usnadnění přístupu a práce s těmito publikacemi. Digitalizace slouží mimo jiné k dlouhodobému uchování kulturního dědictví. V současné době vzniká na celém světě nespočet digitalizovaných dokumentů v různých formátech a jejich počet se každou chvílí zvětšuje. Digitalizované dokumenty nepochybně patří mezi big data.

S digitalizací je velmi úzce spjata metoda OCR neboli optické rozpoznávání znaků (Optical Character Recognition), která umí v nascanovaných dokumentech rozpoznat jednotlivé znaky a převést tak obrazovou předlohu do textové podoby. Při využití metody OCR na digitalizované dokumenty získáváme další velmi užitečné obsáhlé soubory dat, díky kterým je umožněno pracovat s digitalizovanými dokumenty jako s jakýmkoli jiným počítačovým textem, kde můžeme provádět fulltextové prohledávání či kopírovat části textu.

4.3 Šedá literatura

Jednou z oblastí, kde je produkováno velké množství dat, je šedá literatura. Tento pojem označuje veškerou literaturu vytvářenou na všech úrovních vládních, akademických, obchodních a průmyslových institucí, jak v elektronické tak v tištěné podobě, která nebyla vydána standardním vydavatelským procesem prostřednictvím komerčních vydavatelů a není distribuována do běžné prodejní sítě. Šedá literatura je vydávána institucemi, mezi jejichž hlavní činnosti nepatří vydavatelství. [4] Zahrnuty jsou zde vysokoškolské kvalifikační práce, sborníky z různých konferencí, firemní literatura, výzkumné a technické zprávy, webové stránky, blogy a podobně. Díky absenci často zdlouhavého vydavatelského procesu bývá šedá literatura mnohdy aktuálnější než například vydávané články v časopisech.

Místem, kde je šedá literatura shromažďována bývají obvykle knihovny, včetně těch akademických. Knihovny například pomáhají vytvářet v rámci digitálních a digitalizovaných fondů šedé literatury aktivní využití rezervoárů již zmíněných kvalifikačních prací (bakalářských, diplomových, rigorózních, disertačních i habilitačních). Jde o velmi cenné zdroje informací i z pohledu velkých dat, které jsou nejen důležitou formou evidence těchto šedých dokumentů, ale zároveň aktivním a pasivním fondem vědeckých dat a poznatků, který bude využit i do budoucnosti pomocí speciálních metod vytěžování znalostí z dat (data mining).

4.4 Vyhledávání

Vedle správy informačních zdrojů je třeba dostupný fond také zpřístupnit uživatelům knihovny, a to co možná nejpříznivější cestou. K tomu slouží různé

nástroje pro vyhledávání, ať už je to knihovní katalog, repozitáře nebo seznamy elektronických zdrojů.

V současné době čím dál více knihoven zavádí používání tzv. discovery systémů, které podobně jako vyhledávač Google umožňují jednoduché centralizované vyhledávání. Prostřednictvím jediného uživatelského rozhraní mají tak uživatelé možnost prohledávat a přistupovat ke všem typům informačních zdrojů, které jejich knihovna nabízí. Přechodem k discovery systémům je postupně upouštěno od dříve velmi rozšířených paralelních vyhledávačů, které využívaly tzv. federativního vyhledávání, kdy veškerá data zůstávají uložena ve vzdálených systémech, tedy v knihovním katalogu, repozitářích, v případě e-zdrojů u jednotlivých vydavatelů. Při paralelním vyhledávání prochází vyhledávač jednotlivé informační zdroje, zadává zde požadovaný dotaz a výsledky hledání postupně sbírá a uživateli zobrazuje na jednom místě. Tento způsob vyhledávání je poměrně přesný, avšak velmi časově náročný. Na rozdíl od paralelních vyhledávačů si discovery systémy vytváří svůj vlastní centrální index, který je jakousi centrální fyzickou databází popisných dat, analogicky fyzickému souhrnnému katalogu. Při zadání dotazu do discovery systému není vyhledávání zahájeno v každém z dostupných zdrojů zvlášť, ale přímo v centrálním indexu, který by měl obsahovat všechna potřebná data. Vyhledávání prostřednictvím discovery je tak velmi rychlé, může však obsahovat různé nepřesnosti, a to především z důvodu závislosti na aktuálnosti metadat uložených v centrálním indexu. Udržet aktuální tak velké množství dat, která lze jednoznačně řadit mezi big data, není triviální záležitostí.

5. Využitelnost big data v knihovnách pro podporu vědy a výzkumu

Většina v dnešní době dostupných dat je uložena v nestrukturované podobě, je tak velmi obtížné je hromadně zpracovávat. Velkou výhodou dat, se kterými se obvykle setkáváme v knihovnách, je jejich alespoň částečná strukturovanost. Knihovníci mají ve zvyku vytvářet pro veškeré v knihovně dostupné informační zdroje některý z typů metadatového popisu. Knihovnická komunita patří také mezi uživatele technologií sémantického webu, které se snaží stále aktivně rozvíjet. V následujícím textu je rozebráno několik oblastí, ve kterých mohou být knihovny užitečné pro oblast vědy a výzkumu díky zpracování v knihovnách dostupných big data blíže představených v předchozí kapitole.

5.1 Bibliometrické analýzy a scientometrie

Zkoumáním aktuálně dostupných informačních zdrojů, článků v časopisech či vydávaných knih, lze sledovat určité trendy v oboru. V publikační činnosti se odráží, na čem vědci pracují, o co se zajímají a jakým směrem se ubírají. Jedná se též o jednu z možností, jak zjistit, že někdo jiný na světě pracuje na pro nás zajímavém tématu. Analýzami publikační činnosti se zabývá obor bibliometrie, jehož hlavním úkolem je aplikace matematiky a statistických metod na knihy a další komunikační média. [5]

Aplikace bibliometrie do oblasti vědy je pak úkolem oboru scientometrie, který se soustředí na hodnocení vědy a výzkumu podle publikační činnosti

pracovníků dané instituce. [6] Díky bibliometrickým analýzám, tj. obvykle kvantitativním hodnocením publikací a jejich citací, lze provádět porovnání vybraných zemí, celých institucí nebo jednotlivých výzkumných pracovišť. Dále můžeme srovnávat jednotlivé obory mezi sebou. Na základě různých bibliometrických analýz bývají také rozdělovány finanční prostředky ve vědě. Tento způsob hodnocení vědy a výzkumu je ve světě poměrně rozšířen, v České republice se na něm momentálně pracuje v rámci projektu IPN Metodika, který si klade za cíl vytvořit návrh nového systému hodnocení a financování výzkumu, vývoje a inovací. [7]

Knihovny mohou být v tomto směru užitečné, neboť mají přístup nejen k potřebným datům ale také k nástrojům, pomocí nichž se dají tato data zpracovávat.

5.2 Cílená akvizice - zajištění aktuálnosti a relevance fondu

Mezi hlavní úkoly knihovny patří udržování aktuálnosti a co možná největší relevance knihovního fondu, a to jak tištěného tak elektronického. Pro tyto účely je vedle udržování si přehledu v daném oboru třeba sledovat různé druhy uživatelského chování.

Jednou z možností se značnou výpovědní hodnotou je sledování logů z vyhledávání prostřednictvím discovery systému a různých knihovních katalogů. Na základě vyhledávaných dotazů můžeme vypozorovat, o jaká témata mají uživatelé zájem nebo jaké zdroje jim v knihovně chybí (vyhledávání, které nevrátilo žádné výsledky).

Vedle vyhledávání jsou pro knihovny důležité informace, jaké dostupné zdroje byly doopravdy využity. Tyto informace získávají knihovny pro tištěné tituly na základě informací o výpůjčkách, ať už knihovních či meziknihovních. Elektronické zdroje nabízí možnosti sledování stahování plných textů. Pozorováním růstu a poklesu zájmu o dokumenty v určitých oborech pak můžeme vysledovat budoucí směr vývoje daného oboru. Tyto informace jsou důležité mimo jiné pro zkoumání procesu zastarávání informací, který tvoří základ obsahové prověrky fondu a vede k vyřazování dokumentů, jejichž existenci již neospravedlňuje ani informační ani historická hodnota.

Efektivní využívanost nakoupených informačních zdrojů v akademických knihovnách se dále odráží v citovanosti. U zdrojů v publikacích jednotlivých autorů, které autoři citovali, se dá předpokládat, že je autoři také četli. Pořízení takových zdrojů můžeme považovat za dobře investované peníze. Na základě citací lze spočítat návratnost investice do jednotlivých informačních zdrojů.

5.3 Služby pro vědce

Kromě zpřístupňování různých typů informačních zdrojů mohou knihovny poskytovat i různé typy užitečných služeb. Díky přehledu o aktuálně vydávaných publikacích jsou knihovníci schopni poskytovat cenné informace ohledně toho, kde a jak publikovat. S tím souvisí například současný trend open accessu, se kterým se zatím ne všichni vědci sžili, ale knihovníci ho již obvykle považují za běžnou záležitost.

Za velmi užitečné lze považovat rešeršní služby, díky nimž pomáhají knihovníci vědcům zorientovat se v určité problematice za pomoci již publikovaných informací na dané téma.

Další služby lze pak vytvářet například na základě analýzy uživatelského chování, pohybu na webu, vyhledávání atd. Cílem je uzpůsobit veškerou nabídku služeb na míru uživatelům konkrétní knihovny, v případě vědeckých a akademických knihoven vědeckým a pedagogickým pracovníkům a studentům příslušné instituce.

5.4 Spojení s tématikou rešeršních služeb a rešeršních strategií

Pro práci s oblastí big data je spojena také schopnost a umění využít tohoto velkého objemu dat ze strany uživatele, ale také napomoci lepší interpretaci, která napomůže pochopení, ale i vidění souvislostí, které umožní efektivnější přijetí rozhodnutí. Proto se mezi rešeršní nástroje dostávají ve stále větší míře volně přístupné nebo komerční produkty, které to umožňují. Je možno vidět souvislost i s technologiemi a postupy tzv. vytěžování dat (data mining). Je možno uvést celou řadu příkladů na spojení této oblasti s prostředky typu data mining a s rešeršními strategiemi. Pracují s nimi některá databázová centra sloužící výzkumu a vývoji včetně medicínského. V oblasti vědy a výzkumu, vědeckotechnických informací, je takovým subjektem a silným hráčem databázová síť STN International (*http://www.stn-international.de*), kde se mj. nachází jedny z nejprestižnějších databází pro lékařské, farmaceutické a biochemické obory. Informační služba STN ANAVIST pak s daty převedenými do přehledových zpráv (reportů) umožňuje pracovat v zajímavých souvislostech. Jde o nástroj interaktivní analýzy a vizualizačního software, který nabízí celou řadu způsobů, jak analyzovat výsledky hledání z odborné literatury, ale také patentů, stejně tak nabízí vizualizaci pro pochopení trendů ve výzkumu a inovacích. Díky práci s daty a dalšími nadstavbovými nástroji (jde o komerční službu a poměrně drahou pro uživatele) je možno analyzovat patentovou aktivitu, provádět činnosti pro účely tzv. competitive intelligence, ale také stanovit výzkumné trendy a směry, jejich sílu či úpadek, podporovat strategické plánování.

Jiným příkladem může být použití prostředku IBM I2, kdy je možno díky vizualizačním technikám analyzovat velké množství dat a přijatelnou formou je nabídnout k pochopení a využití ze strany uživatele. Takové a podobné nástroje jsou běžně k dispozici v podnikové nebo státní sféře, ale je otázkou času, kdy se takové nástroje dostanou do knihoven [3] včetně lékařských. Příkladem práce s velkým množstvím dat do budoucna můžeme očekávat i u tématiky EBM (evidence-based medicine). Takové nástroje jsou k dispozici již dnes, ale nepracují ještě s takovým množstvím dat, které bychom označovali v současném chápání pojmu „big data". Do problematiky big data vstupují pak prostředky umělé inteligence a pokročilých znalostních systémů. Speciální kapitolou využití jsou také služby sociálních sítí, ať už pro účely marketingu nebo pro sledování trendů chování jejich uživatelů.

6. Závěr

Není pochyb o tom, že knihovny disponují množství různých big data, a to především v oblastech publikační činnosti, digitalizace a vyhledávání v knihovních katalozích. Tato data mohou být potenciálně využitelná pro oblast vědy a výzkumu prostřednictvím bibliometrických analýz a s nimi související scientometrie a cílené akvizice zajišťující čtenářům knihovny pro ně relevantní a aktuální fond informačních zdrojů. Dále mohou být na základě znalostí knihoven v oblasti publikování a analýz uživatelského chování vytvářeny služby uživatelům knihovny na míru.

Literatura

[1.] Hurwitz, J., Nugent, A., Halper, F. and Kaufman, M. Big data For Dummies, Wiley: 2013.

[2.] Long, M. P. and Schonfeld, R. C. Ithaka S+R US Library Survey 2013, 2014.

[3.] Papík, R. and Giannetti, G. Knihovny mají nástroje, o nichž se firmám ani nesní, internet je nemůže nahradit. Lupa.cz [Online] 2012, http://www.lupa.cz/clanky/knihovny-maji-nastroje-o-nichz-se-firmam-ani-nesni-internet-je-nemuze-nahradit/.

[4.] GL'97 Proceedings: Perspectives on the Design and Transfer of Scientific and Technical Information, Third International Conference on Grey Literature, Luxembourg, 1997; Farace, D. J.; TransAtlantic: Amsterdam, 1998.

[5.] Plitchard, A., Statistical Bibliography or Bibliometrics. Journal of Documentation, 1969. 25(4): p. 348–349.

[6.] Diodato, V., Dictionary of Bibliometrics. 2012: Routledge.

[7.] MŠMT IPN Metodika, http://metodika.reformy-msmt.cz/.

Kontakty:

Ing. Anna Motejlková
1. Národní technická knihovna
2. Centrum informačních služeb VŠCHT Praha
3. Ústav informačních studií a knihovnictví FF UK
e-mail: anna.motejlkova@techlib.cz
tel.: +420 232 002 572

doc. PhDr. Richard Papík, Ph.D.
1. Ústav vědeckých informací 1. lékařské fakulty UK
2. Ústav informačních studií a knihovnictví FF UK
e-mail: papikr@cuni.cz
tel.: +420 251 080 205