EBIB Biuletyn EBIB, nr 8 (135)/2012,
e-nauka — wyzwania dla bibliotek akademickich
Dział artykuły tematyczne

Bożena Bednarek-Michalska
Biblioteka Główna
Uniwersytet Mikołaja Kopernika

Repozytoria surowych danych — dlaczego biblioteki powinny je znać?

Streszczenie: *Autorka stara się przedstawić nowe formy gromadzenia i opracowywania surowych danych powstających w wynikach procesów badawczych, które są dość znane w świecie, ale nie praktykowane w Polsce. Opisuje czym są takie dane i jak się je gromadzi oraz upowszechnia. Wylicza także repozytoria takich danych oraz informuje o inicjatywach europejskich.*

Słowa kluczowe: *surowe dane badawcze, repozytoria danych, model open access,*

Idea, ruch i model publikowania naukowego open access są znane polskim bibliotekarzom od lat. Wiedzą oni, że jest związana z upowszechnianiem zasobów nauki, wiedzą także, że wiązała się z publikacjami naukowymi. Ale w dzisiejszym świecie technologii informacyjnych wszystko tak szybko się zmienia, że ci, którzy śledzą nowinki w tym zakresie wiedzą już więcej. Zmienia się cały system prowadzenia badań, edukacji i upowszechniania informacji. Dziś nie mówi się tylko o otwieraniu zasobów nauki, ale o otwartych laboratoriach, procesach badawczych, konferencjach, debatach i narzędziach im towarzyszących, a także otoczeniu prawnym.

Jednym z tych nowych aspektów otwartej nauki jest zbieranie otwartych danych. Powstają repozytoria danych badawczych, danych surowych, którym warto się przyjrzeć, bo są różnorodne i trudne do unifikacji oraz gromadzenia. Tworzą je wielkie konsorcja instytucji naukowych, centra badawcze, czasem biblioteki uniwersyteckie. Jaka będzie ich przyszłość? Czas pokaże, ale sądzę, że może to być jedno z nowych zadań dla bibliotekarzy.

Czym są otwarte dane badawcze?

Nauka i badania opierają się przede wszystkim na gromadzeniu, analizie danych, publikacji, powtórnej analizie krytycznej i ponownym wykorzystaniu danych. Obecny system publikacji nie sprzyja upowszechnianiu danych, dostęp do nich często jest bardzo utrudniony, bo nie wszystkie są publikowane, nie mają żadnych opisów, adnotacji prawnych itp. Naukowcy także niechętnie je udostępniają.

Dostępność danych jest także ograniczana przez inne czynniki. W państwach Unii Europejskiej zmiany w systemie prawa własności intelektualnej umożliwiają ochronę baz danych na zasadzie podobnej do ochrony twórczości przez prawo autorskie. Fragmenty danych również mogą być chronione prawami autorskimi (w niektórych krajach chroni się np. opisy bibliograficzne).

Część wydawców traktuje naukowe bazy danych jako źródło prywatnego zysku, ograniczając dostęp do nich za pomocą kontraktów prawnych (np. nie wolno wysyłać danych przez wypożyczalnię międzybiblioteczną) oraz utrudnień technicznych. Brak

odpowiednich standardów publikacji surowych danych powoduje, że ich zbieranie, przetwarzanie i agregowanie jest utrudnione. Wreszcie w niektórych przypadkach nie są jasne warunki prawne dostępu do danych — nawet wtedy, gdy w intencji twórców bazy ma to być dostęp otwarty.

Zapewnienie pełnej dostępności danych wymaga więc zniesienia trzech rodzajów barier: ekonomicznych, prawnych i technicznych, ponieważ brak publicznej dostępności do danych pociąga za sobą szereg skutków ubocznych:

- wyższe koszty prowadzenia badań,
- ograniczenie poziomu badań ze względu na uciążliwość pozyskiwania danych,
- ograniczenie opartej na dostępności danych naukowych innowacyjności w gospodarce,
- ograniczenie współpracy naukowej, szkoleń i edukacji,
- gorsza jakość danych, które nie podlegają publicznej weryfikacji,
- wzrost barier cywilizacyjnych między państwami rozwiniętymi i rozwijającymi się.

Otwarte dane — Panton Principles

W roku 2009 naukowcy Peter Murray-Rust, Cameron Neylon, Rufus Pollock i John Wilbanks spisali w Cambridge (Wielka Brytania) kilka zasad odnoszących się do prawnej otwartości danych badawczych. Zasady te zostały potem dopracowane przez członków grupy roboczej Open Knowledge Foundation Working Group on Open Data in Science i oficjalnie zaanonsowane w lutym 2010 r. Są one dziś znane pod nazwą *Panton Principles* (http://pantonprinciples.org/[1]).

Formalnie zaleca się, by wszystkie dane i publikacje wytworzone za publiczne pieniądze od razu przechodziły do domeny publicznej, były dobrem wspólnym. Można wykorzystać do tego licencje Public Domain Dedication lub licencje Creative Commons „Zero". Jeśli nie uda się wypełnić takiego warunku, to wylicza się inne możliwości, które powinny być spełnione:

1. Udostępnionym danym powinno towarzyszyć klarowne oświadczenie dotyczące woli i oczekiwań co do ponownego użycia zarówno pojedynczych rekordów, części, jak i całości danych. Takie oświadczenie powinno być precyzyjne, nieodwołalne i oparte na wybranej formule licencji lub całkowitym zwolnieniu z ochrony. Kiedy udostępnia się dane, powinno się formalnie określić zasady ich wykorzystania.
2. Wiele powszechnie uznanych licencji nie jest odpowiednich dla publikowania otwartych, surowych danych lub zbiorów danych. Licencje, które są stworzone i dostosowane do stosowania w odniesieniu do zasobów sieci, takie jak: Creative Commons, GFDL, GPL, BSD itp., nie są odpowiednie dla danych, a ich użycie jest odradzane. Zaleca się używanie licencji odpowiednich dla danych, które nazywają się Conformant Data Licenses (http://opendefinition.org/licenses/#Data).
3. Nie powinno się stosować licencji, które ograniczają ponowne handlowe wykorzystanie lub ograniczają produkcję dzieł pochodnych, wyłączających

[1] Wszystkie odwołania do stron internetowych zawierają dane aktualne w dniu 7 sierpnia 2012 r.

stosowanie ich do określonych celów, konkretnych osób czy organizacji — to jest zdecydowanie odradzane. Licencje takie uniemożliwiają skuteczne ich zintegrowanie i ponowne zastosowanie w działalności gospodarczej, gdzie mogłyby zostać wykorzystane z pożytkiem.

Bariery ekonomiczne można znieść, obwarowując finansowanie publiczne zasadą „zwrotu podatnikom poniesionych nakładów", a technologiczne wykluczyć przez stosowanie otwartego oprogramowania, którego jest coraz więcej.

Listy repozytoriów

W 2009 r. w Londynie podczas spotkania British Library, the Technical Information Center of Denmark, TU Delft Library, the National Research Council's Canada Institute for Scientific and Technical Information (NRC-CISTI), California Digital Library, Purdue University oraz German National Library of Science and Technology postanowiono stworzyć stowarzyszenie, które zajmowałoby się upowszechnianiem informacji o danych surowych (*datasets*). W związku z tym, że w Internecie zaczęły pojawiać się repozytoria, bazy surowych danych, powstających w wyniku prowadzenia różnorodnych badań, na początek stworzono spis — listę repozytoriów tych danych — DataCite (http://datacite.org/repolist), by ułatwić naukowcom poruszanie się w gąszczu powstającej nowej informacji naukowej. Następnie stowarzyszenie udostępniło inne usługi, między innymi zaczęło tworzyć i popularyzować standardy oraz przydzielać identyfikatory danym. Główne cele organizacji to:

- wspieranie naukowców, pomoc w odnajdywaniu, identyfikacji, cytowaniu i wymianie wiarygodnych zbiorów danych badawczych,
- wsparcie centrów danych przez dostarczanie im trwałych identyfikatorów dla zbiorów danych, standardów ważnych przy upowszechnianiu danych i wzmocnienie procesów pracy,
- wspieranie wydawców czasopism, umożliwiając im dołączanie do artykułów naukowych danych źródłowych.

Na liście DataCite znajduje się wiele ciekawych zasobów, między innymi:

1. Australian Antarctic Data Centre,
2. British Atmospheric Data Centre,
3. Canadian Space Science Data Portal,
4. Center for International Earth Science Information Network,
5. Chemical Database Service,
6. MorphoBank,
7. Paleobiology Database,
8. SeaDataNet,
9. Statistics New Zealand Data Archive,
10. UK Air Quality Archive,
11. UK Data Archive,
12. World Data Centre for Glaciology,
13. GenBank,
14. Protein Data Bank.

Inne listy repozytoriów danych surowych:

1. DataBib (http://databib.org/) — bibliografia adnotowana tworzona przez Institute of Museum and Library Services.
2. Global Change Master Directory (http://gcmd.nasa.gov/KeywordSearch/Home.do) — serwis informacyjny tworzony przez Earth Sciences Directorate w NASA.
3. Open Access Directory Data Repositories (http://oad.simmons.edu/oadwiki/Data_repositories) — lista tworzona od 2008 r. przez studentów i pracowników Simmons College.
4. Public Data Sets on Amazon Web Services (http://aws.amazon.com/datasets) — niekomercyjny serwis Amazona, związany z informowaniem o naukowych bazach danych surowych.

Interesującą usługę *Eksplorator danych publicznych* wprowadziła w roku 2010 na rynek firma Google (http://support.google.com/publicdata/). Jest to wyszukiwarka, która umożliwia przeglądanie, wizualizację i przekazywanie danych publicznych, takich jak: wykresy, mapy, dane statystyczne. Na razie udostępniane są tylko z Banku Światowego, ale firma zapowiada rozwój zasobu. Linki do danych są automatycznie aktualizowane.

Przykładowe repozytoria

W Internecie można znaleźć już repozytoria otwartych danych, nie tylko o charakterze naukowym, np. Bank Światowy udostępnia od niedawna w sposób otwarty swoje dane na stronach http://data.worldbank.org/. Są tam różne katalogi i bazy surowych danych, tabel, wykresów, raportów i inne. Można je pobierać i przetwarzać dowolnie: do celów informacyjnych czy badawczych, komercyjnych i non-profit. Warunki tego użycia są jasno określone w zasadach umieszczonych przy bazach (http://data.worldbank.org/summary-terms-of-use).

Międzynarodowa organizacja naukowa International Council for Science stworzyła w roku 2010 portal do danych surowych World Data System (WDS, http://www.icsu-wds.org/services/data-portal), w którym można znaleźć otwarte dane biologiczne, fizyczne, ekonomiczne, geograficzne i inne. Danych dostarczają międzynarodowe instytucje naukowe i członkowie konsorcjum tworzący bazę, w tej chwili w takich zakresach jak: astronomy, biodiversity, cartography, and geodesy, climate, Earth sciences, ecology, general, geodesy and geophysics, geoinformatics and sustainable development, geomagnetism, glaciology, laser ranging, marine environmental sciences, marine geology and geophysics, meteorology, oceanography, remotely sensed data, renewable resources and environment, rockets and satellites, seismology, soils, Solar activity, Solar terrestrial physics, solid Earth physics, space science, sunspot index. Zasady udostępniania dostępne on-line (http://icsu-wds.org/organization/data-policy) są zgodne z otwartą polityką znanej międzynarodowej organizacji The Group on Earth Observations (GEO), gromadzącej wszelkie dane o kuli ziemskiej na nowym Portalu GEO (http://www.geoportal.org/).

Przykładowe metadane z WDS:

Wolff, Katrin (2011): Sub-annual dust concentration and size distribution data from ice core NGRIP sections corresponding to the early Holocene, Allerod IS, LGM, late glacial and DO-7. *Alfred Wegener Institute for Polar and Marine Research, Bremerhaven*, Unpublished dataset #770167.

Anonymous (2004): Biological, chemical, and physical data from CTD/XCTD from five Japanese R/Vs in the North Pacific Ocean from January to December 2002 (NODC Accession 0001334).

Data Center: NODC: National Oceanographic Data Center

Summary: Biological, chemical, and physical data from CTD/XCTD from five Japanese R/Vs in the North Pacific Ocean from January to December 2002 (NODC Accession 0001334)

Parameters: NUTRIENTS; PRIMARY PRODUCTIVITY; SALINITY; TEMPERATURE; ZOOPLANKTON

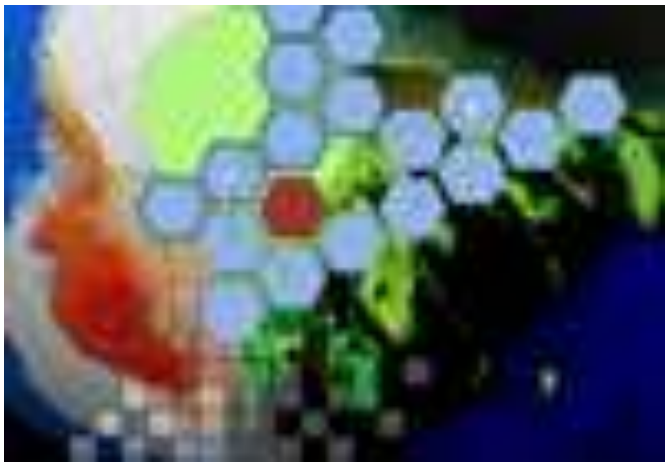
Do opisów bibliograficznych w takich bazach dołącza się wykresy, fotografie, schematy, pomiary itp.

Innym przykładem jest multidyscyplinarne repozytorium Edinburgh DataShare (http://datashare.is.ed.ac.uk/) tworzone przez University of Edinburgh, a urzymywane i monitorowane przez bibliotekarzy. Pracownicy naukowi, którzy tworzą, gromadzą dane związane z późniejszymi publikacjami naukowymi czy raportami są zachęcani do ich gromadzenia, archiwizowania w bazie i dzielenia się z innymi. Wszystkie dane mają unikalne identyfikatory, opatrzone są dokładnymi metadanymi i mają jasno określony status prawny.

Poniżej podaję przykładowe skrócone opisy, które oczywiście można rozwinąć i przy których znajdują się pliki do załadowania w różnych formatach w zależności od tego, jak dane były gromadzone i czym przetwarzane, są one często spakowane (.zip) ze względu na swoją wielkość:

1. Survey of Scottish Witchcraft, 1563-1736 Goodare, Julian; Yeoman, Louise; Martin, Lauren; Miller, Joyce (University of Edinburgh. School of History, Classics and Archaeology., 2010-08-18).
2. Database of Dedications to Saints in Medieval Scotland John Davies; Eila Williamson; Steve Boardman (University of Edinburgh, School of History, Classics and Archaeology, 2008-09-05).

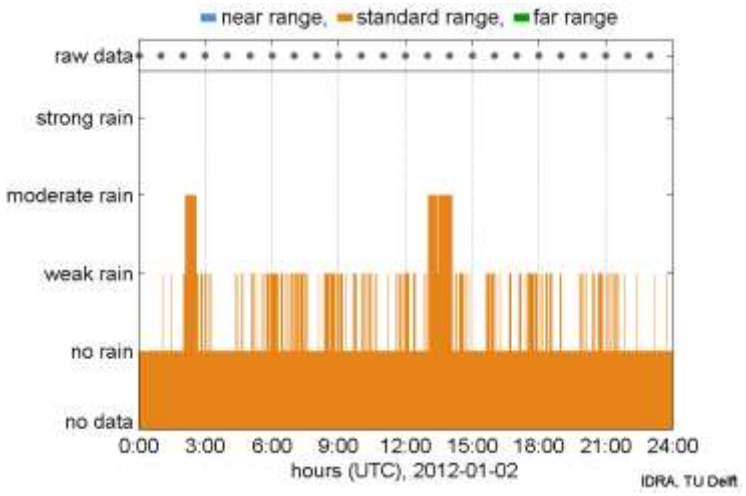
Bardzo interesujące jest rozwiązanie narodowe Holandii, która ma swój wielodziedzinowy portal naukowy DANS (Data Archiving and Networked Services, http://www.dans.knaw.nl/en), udostępniający dane badawcze w systemie EASY na różnych zasadach w zależności od rodzaju danych. Wiele z nich jest otwartych, ale nie wszystkie, ponieważ organizatorzy zostawiają decyzję co do licencjonowania właścicielowi danych. Zaleca się, by dane były udostępnione w modelu open access, ale o ile ich właściciel chce, może wskazać warunki, na jakich dane mogą być wykorzystane. Więcej informacji o zasadach znajduje się na stronach portalu: (http://www.dans.knaw.nl/en/content/dans-licence-agreement-deposited-data).



1. Cyfrowe geodane.
Źródło: Digital preservation geodata. W: DANS: Data Archiving and Networked Services [on-line]. [Dostęp 17.10.2012]. Dostępny w World Wide Web: http://www.dans.knaw.nl/en/content/categorieen/nieuws/digital-preservation-geodata.

DANS tworzony jest przez Koninklijke Nederlandse Akademie van Wetenschappen, KNAW oraz De Nederlandse Organisatie voor Wetenschappelijk Onderzoek, które podpisują umowy z innymi instytucjami, np. Kadaster czy organizacjami rządowymi gromadzącymi dane potrzebne do badań.

DANS tworzy także narodowy portal naukowy NARCIS, który gromadzi wszystkie informacje naukowe, nie tylko o publikacjach, ludziach, raportach, ale i danych surowych. Ambicją Holendrów jest inkorporować do tego systemu wszystkie naukowe zasoby kraju. Oto przykładowy wykres pokazujący pogodę w konkretnym dniu z miejscowości Cabauw w Holandii:



2. Pogoda w dniu 2 stycznia 2012 r. w Cabauw w Holandii.
Źródło: IDRA weather radar measurements — day 2012-01-02. W: Datasets 3TU.Datacentrum [on-line]. [Dostęp 17.10.2012]. Dostępny w World Wide Web: http://data.3tu.nl/repository/uuid:108b208e-00b8-414f-aebb-995635c0b1a4.

The UK Data Archive (http://www.data-archive.ac.uk/about/archive) jest od 1967 r. opiekunem największej kolekcji danych z zakresu nauk społecznych i ekonomicznych w Wielkiej Brytanii. Jest to nie tylko archiwum danych potrzebnych do badań, ale i centrum wszelakich porad dla tych, którzy gromadzą dane surowe. Dostęp do większości danych jest darmowy i dostępny dla większości, ale trzeba założyć profil instytucji i logować się. Instytucje spoza Wielkiej Brytanii mają dostęp do mniejszej ilości danych.

Przykładowy opis z zasobu:

Title: Polish National Election Study, 2001, (PNES; **Poland**'s Third Transition : Beyond 2001)
Subject Categories: Government and political systems — Politics, Election and campaign studies — Politics
Depositor(s): McManus-Czubinska, C., University of Glasgow. Department of Central and East European Studies
Principal Investigator(s): McManus-Czubinska, C., University of Glasgow. Department of Central and East European Studies, Miller, W.L., University of Glasgow. Department of Politics, Markowski, R., Polish Academy of Sciences. Institute of Political Studies, Wasilewski, J., Polish Academy of Sciences. Institute of Political Studies
Data Collector(s): Centrum Badania Opinii Społecznej (Warsaw, **Poland**).

W Polsce o otwartych danych mówi się na razie w kontekście danych publicznych. Problemem tym zajmuje się Koalicja Otwartej Edukacji, a szczególnie jej partner Centrum Cyfrowe Projekt: Polska, który tworzy serwis poświęcony takim danym (http://otwartedane.pl/). W stopce portalu możemy przeczytać: *O ile nie jest to stwierdzone inaczej, dane publiczne udostępnione w serwisie nie są przedmiotem praw autorskich. Wszystkie inne treści, o ile nie jest to stwierdzone inaczej, są dostępne na licencji* Creative Commons „Uznanie autorstwa 3.0 Unported*". Pewne prawa zastrzeżone na rzecz* Centrum Cyfrowego Projekt: Polska[2]*.*

Zakończenie

Napisałam ten tekst, jak zawsze, z powodów praktycznych, bo mam świadomość, że budowanie repozytoriów danych surowych będzie należało w przyszłości do bibliotekarzy, a nie jest to zadanie łatwe, w każdym razie dużo trudniejsze niż budowanie bibliotek cyfrowych i repozytoriów publikacji. Dane surowe wymagają innych opisów, innych standardów przechowywania, innych zasad gromadzenia. Nie mamy doświadczenia w tym zakresie, ale takie repozytoria uniwersyteckie już są, można je obejrzeć i poczytać o nich: Edinburgh DataShare (http://datashare.is.ed.ac.uk/) czy eCrystals tworzone na University of Southampton (http://ecrystals.chem.soton.ac.uk/).

Ponadto powstał interesujące opracowanie przygotowane przez bibliotekarzy z Pardue University Library, którzy napisali artykuł *Institutional Repository Data Set Collecting: Outcomes of a Research Library Task Force*, jak takie repozytorium powinno powstawać, co należy przewidzieć, przygotować, z kim rozmawiać.

[2] *Otwartedane.pl: beta* [on-line]. Dostęp 17.10.2012. Dostępny w World Wide Web: http://otwartedane.pl/.

Bibliotekarze amerykańscy już od 2006 r. zajmują się debatą na temat danych badawczych, organizują seminaria i warsztaty, które mają pokazać wagę tego problemu i wskazać go jako nowe wyzwanie dla bibliotek akademickich. Nam przyjdzie poczekać około 5–10 lat — jak mniemam. Bibliotekarze i tak będą szybsi od ministerstwa nauki.

Bibliografia:

1. NEWTON, M.P., MILLER, Ch.C., BRACKE, M.S., Librarian roles in institutional repository data set collecting: outcomes of a research library task force, W: *Libraries Research Publications* [on-line]. Purdue University, o1.2011. [Dostęp 17.10.12]. Dostępny w World Wide Web: http://docs.lib.purdue.edu/lib_research/122/.
2. *A report to the National Science Foundation from the ARL Workshop on new collaborative relationships: the role of academic libraries in the digital data universe* [on-line]. Arlington: Association of Research Libraries, 26–27 September 2006 [Dostęp 17.10.12]. Dostępny w World Wide Web: http://www.arl.org/bm~doc/digdatarpt.pdf.

Bednarek-Michalska, B. Repozytoria surowych danych — dlaczego biblioteki powinny je znać?. W: *Biuletyn EBIB* [online] 2012, nr 8 (135), *e-nauka — wyzwania dla bibliotek akademickich* [Dostęp: 20.11.2012] Dostępny w World Wide Web: http://www.nowyebib.info/images/stories/numery/135/135_michalska_.pdf. ISSN 1507-7187.