



Dr Henryk Hollender
Warszawa – UŁaz.

CZY REPOZYTORIA ODMIENIAJĄ NAUKĘ?

Abstract

As we watch contemporary scientific repositories, we find out that they keep remaining the extension to digital libraries, and that there is no coherent terminology, which would permit establishing a difference between the former and the latter. But the largest and most searched repositories of the world often offer also searchable data collections. They seem to be a need of new science, which produces and absorbs data in high numbers. In particular projects, intended to produce a scientific work (being basically a narration leading to a statement), they are not all fully used. So we see them often outside of repositories, namely in data warehouses. There is no direct link between repositories and warehouses, the terminology covering this area is not yet mature, and it is difficult to predict any coordination and bringing together of data-sets, due first of all to their divergent metadata. When we return to repositories of scientific texts, we find how susceptible to aggregation they are; all the business here could end up with one universal service for a country, and in Poland we should look this way at Polish Scientific Bibliography (PBN). Data collections cover material much more difficult to organize, but it is data – not publications -- management which will be the priority in research in the next decades.

Streszczenie

Obserwując współczesne repozytoria naukowe, spostrzegamy, że nadal są one pochodną bibliotek cyfrowych; brak ujednoliconej terminologii, która by pozwalała precyzyjnie odróżnić jedno od drugich. Ale największe i najszerzej wykorzystywane repozytoria świata poza tekstami naukowymi zawierają jeszcze często przeszukiwalne zbiory danych. Wydają się one być potrzebą współczesnej nauki, która wytwarza i pochłania wielkie ilości danych; w poszczególnych przedsięwzięciach, mających na celu wytworzenie tekstu naukowego (narracji z odpowiednią argumentacją, wiodącą do twierdzenia naukowego) nie ulegają one pełnemu wykorzystaniu. Toteż widzimy je coraz częściej także poza repozytoriami, gdzie wchodzi niekiedy w skład hurtowni danych. Związki hurtowni z repozytoriami są umowne, zależą od przyjętej terminologii, trudno tu oczekiwać koordynacji i uwspólnień, głównie ze względu na rozbieżność metadanych. Natomiast świat repozytoriów tekstów naukowych jest podatny na agregację i najprościej byłoby mieć jeden taki uniwersalny serwis w skali kraju; mogłaby stać się nim Polska Bibliografia Naukowa. Zbiory danych zawierają materiał o wiele trudniejszy do organizacji, ale to właśnie gospodarka danymi, nie tekstami będzie priorytetem naukowym w najbliższych dekadach.

Powszechnie znane są przyczyny zainteresowania repozytoriami, a największe z nich w coraz większym stopniu skupiają na sobie uwagę naukowców. Jeśli posłużymy się serwisem Webometrics Ranking Web of Repositories, znajdziemy najbardziej skuteczne i największe repozytoria polskie – w Poznaniu (oczywiście AMUR), Łodzi, Wrocławiu i Krakowie, dostrzegając natychmiast, że wśród nich znajdują się także biblioteki cyfrowe, poczynając od działającej już wiele lat Kujawsko-Pomorskiej BC¹. Są to pełnotekstowe bibliografie piśmiennictwa naukowego, w wariancie bibliotek cyfrowych zmieszanego także ze zdigitalizowanymi dokumentami zabytkowymi. Nikt nie powie, że nie pasują one do repozytoriów, bowiem mogą być traktowane jako materiały źródłowe do badań naukowych.

Repozytoria prowadzone w Polsce otwierają się od poz. 418 rankingu światowego, co nasuwa skojarzenie ze sferą, w jakiej w innym rankingu Webometrics znajdują się polskie uniwersytety. Czy kraj nasz jest w stanie inicjować nowe kierunki w obrębie dokumentowania i udostępniania wyników badań naukowych? W Europie AMUR (Adam Mickiewicz University Repository) jest na pozycji 109. Dobrze wiedzieć, że wyprzedza go kilka serwisów stworzonych w krajach Europy Wschodniej, takich jak Czech Digital Mathematics Library (49), Elektronnaya Biblioteka BGU (Belarusskij Gosudarstvennyj Universitet², Mińsk, 59), Elektronnyj Naukowyj Archiv Naukowo-Technicznej Biblioteki Nacionalnogo Universitetu „Lvivska Politechnika” (98). W drugiej europejskiej dwusetce mamy jeszcze Jekaterynburg (Uralskij Ffederalnyj Universitet), Debreczyn, Brno, ponownie Mińsk (Belarusian National Technical University), Rygę i Kijów. Widać, że w skali światowej ruch tu jest ogromny, a przyjęcie globalnych standardów odegrało rolę ujednolicającą, umożliwiło bowiem szybkie nadrobienie zaległości.

Standardy dotyczą jednak przede wszystkim warunków technicznych; nie przesądzą o zawartości repozytoriów ani o randze trafiających tam tekstów (mierzonej choćby cytowalnością). Nazewnictwo też nie zostało ujednolicone. Przeglądając serwisy odnotowane w Webometrics, zauważamy, że „repozytorium” raz jest nazwą gatunkową, a raz – własną. Popularne nazwy poza (institutional) repository i (digital) library to po prostu publications (z nazwą instytucji), open archive, open bibliography, digital archive, electronic publications, e-prints (server), document server, research portal, Forschungsplattform, e-theses (dla prac na stopień naukowy), research online oraz nazwy cokolwiek literackie: Eldorado, Parade... Nazwy te wskazują niekiedy na znaczącą

¹ <http://repositories.webometrics.info/>. Niniejszy materiał został zebrany na potrzeby prezentacji w ramach XXXIV Konferencji Problemowej Bibliotek Medycznych „Biblioteka Przyszłości”, 29 czerwca – 1 lipca 2016 r. w Białymstoku. W lipcu tegoż roku ukazała się nowa edycja danych Webometrics, która nie została już uwzględniona w opracowaniu. Zmiany są zresztą niewielkie.

² Nazwy w językach zapisywanych cyrylicą nie transliterujemy, poprzestając na intuicyjnej transkrypcji.

odmienność formuł. Wydaje się jednak, że w większości europejskie repozytoria to dawniejsze biblioteki cyfrowe, stworzone dla zrealizowania „cyfryzacji zasobów” (digitalizacja), z funkcją repozytoryjną „dodaną” z czasem do zasobów, które nie generują dużego ruchu w Internecie lub – powiedzmy ostrożniej – które mogą generować taki ruch w zależności od zupełnie innych kryteriów niż zainteresowanie naukowe.

Jeśli spojrzymy na ranking światowy i jego najwyższe rejony, odkryjemy stosunkowo duże „zagęszczenie” repozytoriów o innej nieco formule niż bibliografia pełnotekstowa. Spójrzmy na przykład na Astrophysics Data System (ADS, 4 miejsce), prowadzony przez Smithsonian Astrophysical Observatory z funduszy NASA. Owszem, jego trzon ma charakter bibliografii, obejmujących ponad 12 mln zapisów³. Twórcy ogłaszają jednak, że „[i]n addition to maintaining its bibliographic corpus, the ADS tracks citations and usage of its records to provide advanced discovery and evaluation capabilities. Integrated in its databases, the ADS provides access and pointers to a wealth of external resources, including electronic articles available from publisher’s websites, astronomical object information, data catalogs and data sets hosted by external archives. We currently have links to over 11.3 million records maintained by our collaborators.” Oznacza to, że z serwisu ADS możemy się np. dowiedzieć o konkretnym przedsięwzięciu naukowym. Równocześnie zachowuje się on jak tradycyjna biblioteka cyfrowa, udostępniając bogate zasoby zdigitalizowanych książek i czasopism – zdigitalizowanych wszakże nie jako „dziedzictwo kulturowe”, lecz jako korpus żywej wiedzy, wykorzystywanej w badaniach naukowych.

Danych dostarcza także CERN Document Server, znajdujący się na 9 miejscu miejsce w rankingu światowym⁴. Zaś serwis Deep Blue Biblioteki Uniwersytetu Michigan⁵ (miejsce 13) od razu ogłasza na swoim pierwszym ekranie, że dzieli się na „documents” i „data”, i faktycznie udostępnia raporty z badań oraz bazy danych, wyszukiwalne poprzez ogólne „googlowskie” okno dialogowe.

Jak się wydaje, w nauce w coraz większym stopniu liczą się dane, które nie przybrały jeszcze postaci narracyjnej, właściwej artykulowi czy książce. Gdy już taki tekst powstaje, liczy się bardziej argumentacja prowadząca do konkluzji niż wyniki badawcze, które do takiej właśnie argumentację przywiodły. Wyniki są ukryte w załącznikach, w aneksach, w niepublikowanych raportach roboczych, ogólnie – w dokumentacji nie przeznaczonej dla czytelnika lub przeznaczonej w postaci selektywnej. Jest to oczywiste: przedmiotem komunikacji naukowej zawsze były twierdzenia naukowe; publikowane piśmiennictwo wystarczało, by odtwarzać drogi do tych twierdzeń. Teraz jednak widzimy, że wielkie przedsięwzięcia badawcze pozostawiają po sobie ogromne ilości danych, będących czymś więcej niż empirycznym uzasadnieniem publikowa-

³ <http://adswwww.harvard.edu/>. Dane z adresów repozytoriów notowane 2 czerwca 2016 r.

⁴ <https://cds.cern.ch/>

⁵ <https://deepblue.lib.umich.edu/>

nych twierdzeń. Dane te mogą prowadzić do innych twierdzeń, są jednak na tyle bogate i zróżnicowane, że zespoły badawcze nie kontrolują ich wykorzystania.

Ale to także komputeryzacja działalności badawczej (aparatura), zespołowy charakter przedsięwzięć badawczych oraz nieograniczone możliwości archiwizowania danych bez względu na poziom ich bieżącej przydatności spowodowały, że współczesna nauka wytwarza dane na potęgę. W niektórych kierunkach tylko bardzo duża liczba danych może prowadzić do jakichkolwiek wniosków, liczy się już więc „pospolite ruszenie” przy ich zbieraniu – wkład amatorów jako tzw. nauka obywatelska, co w biznesie nazywamy *crowdsourcing* i co de facto stanowi wykorzystanie mechanizmów, które dojrzewały wraz z mediami społecznościowymi. Do gromadzenia danych wykorzystujemy nawet umarłych, wzrasta bowiem znaczenie obserwacji czy pomiarów, zanotowanych w przeszłości, zwłaszcza – zmiennych, gromadzonych przez długi okres (klimatologia, sejsmologia, epidemiologia, demografia, ekonomia, archeologia itp.). Nie wytworzy się już tych danych od nowa, można je jednak odmiennie zagregować, co jest wkroczeniem na drogę ich nowego wykorzystania.

Wielka nauka w XXI w. jest coraz bardziej „datochłonna”. Badania przełomowe to właściwie niemal zawsze badania interdyscyplinarne, gdzie dużej liczbie danych pochodzących z różnych obszarów rzeczywistości towarzyszą innowacyjne metody zestawiania ich ze sobą.

Wydaje się zatem, że repozytoria nie tylko wychodzą poza swoje tradycyjne funkcje bibliograficzne i wzbogacają się o dane, ale także – że rodzą się nowe formuły gromadzenia, przechowywania i udostępniania danych, które w ogóle pozostawiają na boku przedsięwzięcia repozytoryjne. Zdaniem Interdyscyplinarnego Centrum Modelowania Matematycznego i Komputerowego UW, „odповідzią na wyzwania naukowe współczesnego świata” jest nowo założony OCEAN – Otwarte Centrum Danych i ich Analiz⁶. Analiza, jak rozumiemy, to jeszcze nie tekst z twierdzeniami naukowymi, ale już takie uporządkowanie danych, by mogły prowadzić ku wynikowi. ICM zajmuje się także danymi w zakresie humanistyki.

Bibliotekarzowi może się tu wydawać, że staje przed nowym zjawiskiem, ale przecież on już je zna. Instytucje od dawna po prostu „wieszają w internecie” swoje wytwory; faktycznie każda placówka badawcza ma jakieś. Może zakładać repozytorium, może po prostu udostępniać pliki pojedynczo; wielu twórców, a nawet użytkowników nie widzi różnicy, docierając do pożądanej informacji intuicyjnie i nie interesując się architekturą wykorzystywanego serwisu. Niejeden użytkownik nazwie go „stronką”, tak jak nazwałby bibliotekę cyfrową czy repozytorium o złożonej strukturze i wielu funkcjonalnościach; tak jak w istocie nazwałby cokolwiek, co ma swój adres w Internecie. „Stronka” może być rozpoznawalna poprzez instytucję, która – tak jak OCEAN

⁶ <http://ocean.icm.edu.pl/>

– organizuje serwis i zapewnia mu obsługę informatyczną, zwłaszcza – infrastrukturę. Można też instytucji nie zauważyć, docierając bezpośrednio (poprzez Google) do interesującego nas pliku.

Systemów takich istnieje wiele – poza bibliotekami cyfrowymi, poza repozytoriami tekstów czy repozytoriami danych. Oto np. Centrum Systemów Informacyjnych Ochrony Zdrowia⁷, które m. in. prowadzi w Centrum Zdrowia Dziecka Krajowy Rejestr Operacji Kardiochirurgicznych⁸, wyróżniony przez Europejski Kongres Kardiochirurgów i Torakochirurgów. Oto np. Leśne Centrum Informacji – platforma informacyjna monitoringu środowiska przyrodniczego, realizowany w ramach programu operacyjnego Innowacyjna Gospodarka⁹. Niektóre usługi Centrum są dostępne w Otwartym Dostępie i mogą być z powodzeniem wykorzystywane przez amatorów. LCI ma swoje moduły w warstwie prezentacji (portal), usług (aplikacje i narzędzia pomocnicze) oraz danych (Centralne Repozytorium Danych i Hurtownia Danych. Nie mowa tu o bazie piśmiennictwa leśnego; tę Instytut prowadzi oddzielnie i od lat.

Jak się zdaje, terminologia i taksonomie nie są w tym zakresie ostatecznie ustalone. Historycznie, źródłem takich serwisów jest bankowość. Zadaniem hurtowni danych jest właśnie przetwarzanie heterogenicznych danych w informację naukową. Przyjmuje się na ogół, że hurtownie są ukierunkowane na realizację celów strategiczno-analitycznych, w przeciwieństwie do transakcyjnych baz danych, które realizują cele operacyjne. W obrębie poszczególnych serwisów, w warstwie danych (przechowywania danych) występuje na ogół repozytorium danych, chociaż termin „repozytorium” stosuje się też wobec zbioru metadanych. W sensie logicznym mamy tutaj do czynienia z pewnym kontinuum, którego przeciwległe krawędzie tworzą biblioteki cyfrowe (tekstów naukowych), repozytoria (wielorakich zastosowań i zawartości) oraz hurtownie danych (*data warehouse*).

Hurtownie te nie są na razie tak zwartymi, wyraźnie wyodrębnionymi serwisami jak biblioteki cyfrowe czy repozytoria – trudno je nazywać, ujmować w zestawieniach i porównywać. Nie mówiąc już o scalaniu, obejmowaniu wspólnymi wyszukiwarkami itp. Serwisy zawierające dane są określane jako datasets, nie jako databases. Próby ich uporządkowania i włączenia do szerszych systemów informacyjno-wyszukiwawczych są na razie bardzo proste. Trudno w ogóle sobie wyobrazić zestaw metadanych, który by mógł sprawdzać się dla danych wytwarzanych na gruncie różnych nauk¹⁰.

Czy nauka polska „widzi” dane, czy dostrzega ich istotność i aktualność związanej z nimi problematyki? *Rozporządzenie ministra nauki i szkolnictwa wyższego w spra-*

⁷ <https://www.csioz.gov.pl/>

⁸ <http://www.krok.org.pl/>

⁹ www.ibles.pl/o-ibl/organizacja/struktura/lesne-centrum-informacji

¹⁰ Por. np. wykaz Scientific Data, prowadzony jako periodyk przez Nature.com (Springer Nature), <http://www.nature.com/sdata/policies/repositories>

wie przyznawania kategorii naukowej jednostkom naukowym, projekt z dn. 6 czerwca 2016¹¹, definiując działalność naukową, wymienia wiele czynności i zadań; niektóre z nich mogą wydać się nam zbyt szczełółowe i niezbyt istotne. Oto lista:

- zatrudnienie przy realizacji badań naukowych lub prac rozwojowych
- publikacje naukowe i monografie naukowe
- działalność innowacyjna (patenty, prawa ochronne, prawa autorskie, prawa do odmian roślin)
- dzieła artystyczne (w tym pełnienie funkcji kuratora wystawy) i dzieła architektoniczne
- uzyskiwanie stopni i tytułów naukowych
- udział w mobilności naukowców.

Ale jest tam też miejsce dla zadań, które – jak wyobrażamy sobie – mogą prowadzić do doceniania repozytoriów danych i hurtowni danych i przyczyniać się do tworzenia ram organizacyjnych dla ich zakładania. Oto one:

- udział w międzynarodowych i krajowych projektach obejmujących badania naukowe lub prace rozwojowe
- udział w tworzeniu strategicznej infrastruktury badawczej
- udział we wdrożeniach i aplikacjach.

Najbliższe lata pokażą, czy problem zarządzania zasobami danych stanie się centralną kwestią nauki. A jeśli do tego dojdzie, to jaki będzie los tradycyjnych repozytoriów tekstów naukowych. Wydaje się jednak, że odpowiedź na to pytanie już znamy. Dane stanowią wyzwanie informatyczne, pełne teksty – już nie. Są przecież wszędzie, a najprościej myśleć o nich jako o materiale wypełniającym jedną centralną bazę. Dla kraju taką bazą powinna stać się Polska Bibliografia Naukowa. Dla codziennego, ogólnego użytku – globalny serwis Google Scholar. A pomiędzy nimi – paleta serwisów specjalistycznych. Dane mogą przejść podobną drogę, lub też pozostać domeną scaleń na poziomie instytucji czy resortu, tak jak w istocie jest w chwili obecnej.

¹¹ <http://legislacja.rcl.gov.pl/docs//506/12286200/12356614/12356615/dokument224647.pdf>