Bożena BEDNAREK-MICHALSKA
Biblioteka Uniwersytecka
Uniwersytet Mikołaja Kopernika w Toruniu

REPOZYTORIA DANYCH BADAWCZYCH DLA HUMANISTYKI

Streszczenie: *W artykule wyjaśniono czym są dane badawcze (datasets, primary research data) i ich repozytoria, kiedy zaczęły powstawać i dlaczego, jaka jest ich rola w odniesieniu do bieżących badań. Przedstawiono ich światowe listy (DataBib, re3data, DataSite) oraz korzyści, której płyną z upowszechniania takich danych w otwartym Internecie. Opisuje zawartość wybranych, najciekawszych archiwów humanistycznych. Analizuje polską sytuację w zakresie tworzenia zbiorów surowych danych w zakresie humanistyki i ich otwartość oraz status prawny zbioru.*

Abstract: Research data repositories for the humanities: *The author describes what research data (datasets, primary research data) and their repositories are, when were built and why, what is their role in relation to the current studies. Presents world registers DataBib and re3data and the benefits from the dissemination of such data in an open Internet. Selects the most interesting databases and describes their contents. Analyzes Polish situation in the creation of raw data in the field of humanities, their openness and the legal status of the content.*
This paper is sponsored by National Science Center (NCN) under grant 2013/11/B/HS2/03048/ Title: "Information Visualization methods in digital knowledge structure and dynamics study".

1. WPROWADZENIE

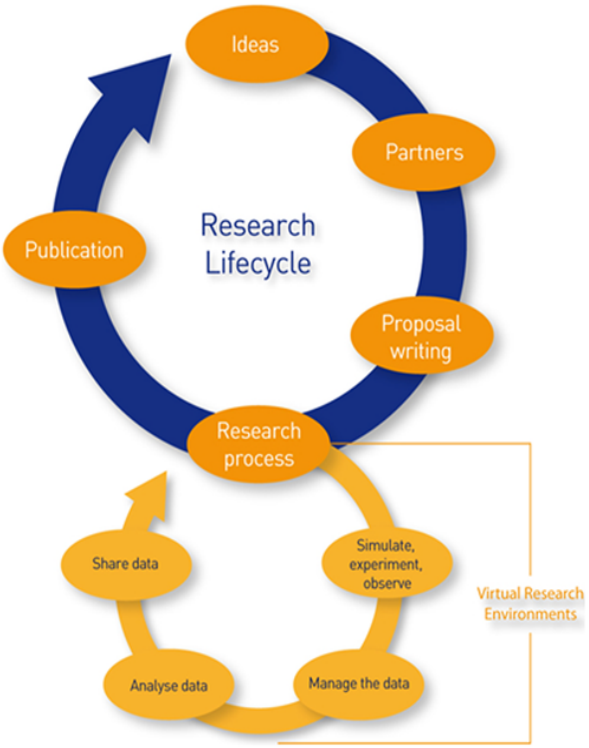
Otwarte dane badawcze to gorący temat w świecie nauki, od kilku lat obserwatorzy środowiska naukowego odnotowują powstawanie interesujących zbiorów danych, które do niedawne były gromadzone w laboratoriach badawczych i szerzej ich nie upowszechniano. Teraz to się zmienia. Gromadzenie i upowszechnianie danych badawczych jest jednym z nowych aspektów otwartej nauki, bardzo trudnym i wymagającym. Dane takie powstają w wielkich konsorcjach instytucji naukowych, centrach badawczych, uczelniach, laboratoriach, czasem bibliotekach uniwersyteckich. Najważniejsza lista światowych repozytoriów danych surowych odnotowuje ich ponad 1200, dwa lata temu było ich o połowę mniej.

2. CZYM SĄ DANE BADAWCZE?

Dane badawcze to informacja, w szczególności zebrane fakty, liczby, które mogą posłużyć badaniom i być traktowane jako podstawa do dalszego wnioskowania, dyskusji lub obliczeń. Przykładowe dane obejmują: statystyki, wyniki eksperymentów, pomiarów, obserwacji wynikających z badań terenowych, ankiety, nagrania wywiadów i zdjęcia. (…)

Otwarte dane badawcze to są takie dane, do których jest powszechny dostęp, które można eksplorować bez przeszkód, wykorzystywać do dalszych badań, powielać, rozpowszechniać bez opłat i innych barier prawnych czy technicznych (EC, 2013b, s. 3).

Dane badawcze powstają w procesach badawczych lub są wyodrębniane z publikacji naukowych. Poniżej można prześledzić cykl życia procesów naukowych i miejsce danych w tym cyklu (Rys. 1). Szerzej zagadnienie to zostanie omówione nieco dalej w artykule.



Rys. 1. Joint Information Systems Committee (JISC), Stages of the research and data lifecycle. Źródło: Tenopir et al., 2011.

3. HISTORIA POWSTAWANIA OTWARTYCH DANYCH

Historia upowszechniania danych badawczych wiąże się z historią upowszechniania danych rządowych i administracyjnych, które były udostępnione znacznie wcześniej chociażby przez portale statystyczne czy meteorologiczne. Otwarty dostęp do danych agend rządowych przyczyniał się do wzrostu ich wykorzystania w procesie badawczym, co nie pozostało obojętnym na stosunek naukowców do upowszechniania danych. Dostrzegli w tym potencjał, który może być zmarnowany, jeśli nie opracuje się mechanizmów wymiany danych między ośrodkami badawczymi i nie zacznie wykorzystywać ich na szerszą skalę.

Komisja Europejska też była sprawą zainteresowana od lat i w roku 2011 opracowała *Open Data Strategy,* która miała na celu uczynić urzędy unijne i narodowe bardziej transparentnymi niż były do tej pory. W strategii zapisano, że trzeba zmienić regulacje prawne dotyczące upowszechniania i ponownego wykorzystania informacji sektora publicznego. Zaplanowane działania Komisji i rządów krajów europejskich mają doprowadzić do tego, by jak najwięcej baz danych administracyjnych zostało otwartych w Internecie (Euroalert, 2011). Niektóre kraje europejskie nie czekając na rekomendacje już upowszechniają dane, np. Wielka Brytania otworzyła narodowe archiwum danych socjologicznych i ekonomicznych nazwane: *UK Data ArChile* (http://www.data-archive.ac.uk/). Bardzo dobre przykłady otwartości przypłynęły także do Europy ze świata. Kilka lat temu Bank Światowy otworzył swoje dane do ponownego wykorzystania na stronach ***World Bank Open Data:*** *free and open access to*

data about development in countries around the globe znajdujących się pod adresem: http://data.worldbank.org/.

W lutym 2013 r. agencje federalne w USA zostały poinformowane przez biuro federalne USA Office of Science and Technology Policy (OSTP), że należy zmaksymalizować dostęp do danych badawczych finansowanych ze środków publicznych (Stebbins, 2013). W czerwcu 2013 r. ministrowie nauki grupy najbogatszych państw świata G8 opublikowali zestaw zasad *G8 Science Ministers Statement* dla nauki w tym otwartych danych naukowych (G8, 2013), które powinny być otwarte w Internecie, by zwiększyć innowacyjność świata.

W związku z wyżej opisanymi działaniami instytucje naukowe zaczęły także upowszechniać swoje dane badawcze z myślą, że mogą się one przydać innym badaczom czy agencjom rządowym, a nawet przedsiębiorcom do ekspertyz, analiz, wprowadzania innowacji czy kontynuowania własnych badań. W 2009 r. 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 przedstawione w lutym 2010 r. Są one dziś znane pod nazwą Panton Principles (Murray et al., 2010).[1]

Komisja Europejska dość szybko zainteresowała się danymi badawczymi, gdyż zauważono, że ich upowszechnianie i dzielenie się nimi może wpłynąć na zwiększenie innowacyjności Europy. Mogą powstawać nowe pola badawcze, te same dane mogą być wykorzystane w różnych aspektach, ich otwartość powoduje, że nie zbiera się dwa razy tych samych danych, co przynosi konkretne korzyści finansowe. Konsekwencją tego zainteresowania było włączenie do programu *Horyzont 2020* testowego rozwiązania w zakresie gromadzenia i otwierania danych badawczych. W celu poznania szczegółowych rozwiązań należy przejrzeć komunikat: *Commission launches pilot to open up publicly funded research data* (EC, 2013a). Stało się podobnie jak wiele lat temu, kiedy testowano modele open access dla publikacji naukowych, co zaowocowało zresztą wdrożeniem konkretnych wymagań grantowych w nowej perspektywie finansowej *Horyzontu 2020.* Szczegóły dotyczące testu można znaleźć w portalu Uwolnij Naukę: http://uwolnijnauke.pl/open-access-w-horyzoncie-2020/, tak teraz zainicjowano test dla danych badawczych (Bednarek-Michalska, 2014). Wydaje się, że ta strategia powolnych acz stanowczych kroków dobrze się w Europie sprawdza.

4. KORZYŚCI WYNIKAJĄCE Z UPOWSZECHNIANIA DANYCH BADAWCZYCH

W Wielkiej Brytanii JISC opublikował raport (McDonald & Kelly, 2012) na temat korzyści płynący z masowej eksploracji danych (ang. *text and data mining*). Wymieniono w nim szereg korzyści, których można się spodziewać z eksploracji w zakresie:

- zwiększenia efektywności badawczej,
- odblokowania ukrytych informacji i stworzenia nowej wiedzy,
- odkrywania nowych horyzontów,
- wzrostu liczby dowodów naukowych,
- usprawnienia procesów badawczych oraz podnoszenia ich jakości,

[1] Szczegóły w moim artykule: Bednarek-Michalska, 2012.

- ooszczędności ekonomicznych i wzrostu wydajności pracy,
- innowacyjnego rozwoju nowych usług,
- powstawania nowych modeli biznesowych,
- tworzenia nowych metod np. leczenia.

Z raportu wynika, że pracując na otwartych danych badawczych rząd Wielkiej Brytanii może zapewnić potencjalne korzyści dla gospodarki brytyjskiej w wysokości do 7,9 mld PLN (Poynder, 2012). Być może z tego powodu rząd Wielkiej Brytanii postanowił finansować przez 5 lat instytut badawczy The Open Data Institute (http://theodi.org/about-us), który ma być katalizatorem zmian i ewolucji otwartych danych badawczych w tym kraju. Instytut stworzony w 2011 r. zatrudnia światowej klasy ekspertów, którzy współpracują, by budować nową kulturę związaną z upowszechnianiem danych, ich eksploracją, gromadzeniem i innowacyjnym podejściem do nauki.

5. SKĄD POCHODZĄ DANE BADAWCZE?

Jak wspomniano wyżej, dane badawcze pozyskiwane są nie tylko w procesach badawczych, ale także z publikacji naukowych. Ten ostatni przypadek dotyczy coraz częstszych praktyk wielkich wydawców piśmiennictwa naukowego, którzy „wydobywają" z artykułów rysunki, wykresy, grafiki i gromadzą je w odrębnych bazach danych, niezależnych od pełnotekstowych. Masowa eksploracja tekstów naukowych staje się także coraz istotniejsza dla samych badaczy, którzy często w ten sposób sprawdzają stan badań w swojej dziedzinie, poszukują podobnych metod, oznaczeń, rysunków, pomiarów. Domagają się oni od wydawców, takich jak Elsevier czy Springer, zapewnienia możliwości technicznych i prawnych na eksplorację danych, co wcale nie jest oczywiste we współczesnym modelu subskrypcyjnym czasopism i baz danych naukowych. Osobą szczególnie zaangażowaną w walkę o tego typu nowe możliwości jest chemik, profesor Peter Murray-Rust.

Przykład wyodrębniania grafik z tekstu można obejrzeć w Public Library of Science w artykule Heinza Pampela i innych (Pampel et al. 2013). Jest powszechną zasadą w tym megaczasopiśmie, że w portalu prezentuje się tekst artykułu oddzielnie, a zamieszczone w nim tabele, ilustracje, wykresy oddzielnie. Te zestawienia i materiały ilustracyjne można pobrać w trzech odmiennych formatach w zależności od potrzeb i wykorzystać we własnym tekście, oczywiście zgodnie z licencją podaną przy źródłowym tekście lub ilustracji. Taka ilustracja ma także przygotowany odrębny opis, oto przykład: *Figure 5. A detailed description of a Research Data Repository.doi:10.1371/journal.pone.0078080.g005.* W publikowanych w ten sposób artykułach wydzielony jest też abstrakt z dokładnym opisem bibliograficznym. Ta fragmentacja tekstu ma swój cel: chodzi o jak najszybsze i najprostsze dotarcie do potrzebnych fragmentów i wykorzystanie ich do innych celów niż pierwotne.

Dane powstałe w procesach badawczych gromadzi się obecnie w różnych bazach danych do tego przystosowanych, opisuje je według określonych zasad i standardów oraz przechowuje na serwerach we własnych instytucjach. Wiele instytucji na świecie oraz wielu uczonych zdecydowało się upowszechnić swoje dane badawcze w Internecie. W tym celu tworzone są otwarte repozytoria, które zwykle są prezentowane ze stron internetowych danej instytucji. Dla ułatwienia ich przeglądu zaczęły powstawać też listy repozytoriów, takie jak DataCite, re3data.org, DataBib. Inicjatywy te w 2015 r. stworzyły alians i połączyły wysiłki, by tworzyć jedną listę: re3data.org.

6. JAK SIĘ UPOWSZECHNIA DANE BADAWCZE?

Dane badawcze upowszechnia się w publikacjach naukowych (książkach i czasopismach) oraz w bazach danych, repozytoriach, archiwach danych badawczych. W tej chwili na świecie istnieje ponad 1200 repozytoriów danych badawczych, a ich wykaz można przeglądać na międzynarodowej liście wyżej cytowanej: http://www.re3data.org/.

Jak pisze na swoim blogu Heinz Pampel (2013), krajobraz repozytoriów danych badawczych nie jest jednorodny. Niektóre inicjatywy, takie jak Data Seal of Approval (DSA) oraz World Data System (WDS), zatrudniają ludzi, którzy pracują nad standaryzacją repozytoriów danych. Wypracowano już procedury certyfikacji i kontroli repozytoriów danych. Jednak normy te nie są jeszcze powszechnie stosowane. Repozytoria danych badawczych są przede wszystkim dziedzinowe. Przechowują pliki w odmiennych formatach, w różnych warunkach, z różnym dostępem do bazy czy ponownego wykorzystania. W wielu przypadkach bywało tak, że trudno było naukowcom znaleźć repozytorium, które byłoby odpowiednie dla przechowywania ich danych lub które dostarczało by im interesującego materiału. Dla przezwyciężenia tego problemu Niemcy rozpoczęli w 2012 r. tworzenie rejestru repozytoriów: *re3data.org* ***Registry of Research Data Repositories*** (grant German Research Foundation - DFG), który zawiera wykaz istniejących repozytoriów.

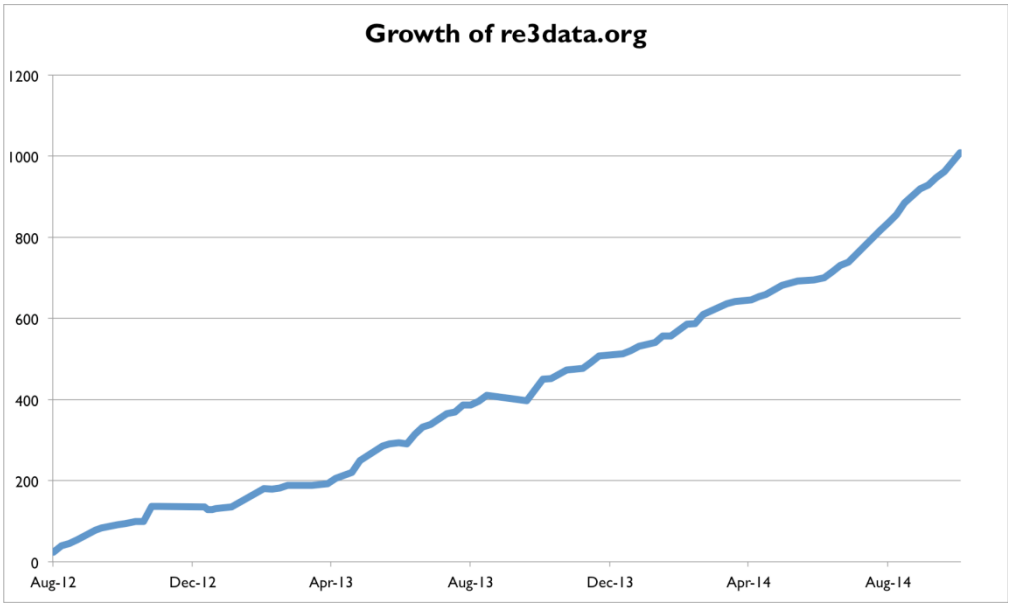
We wrześniu 2013 r. lista *re3data.org* wymieniała 600 zbiorów danych naukowych, dziś jest ich ponad 1200. Wiele z nich szczegółowo opisano także za pomocą słów kluczowych i dziedzin dających możliwość ich wyszukiwania. Rejestr obejmuje zbiory danych ze wszystkich dyscyplin akademickich. Osobą szczególnie zaangażowaną w ten projekt jest Frank Scholze, dyrektor biblioteki akademickiej Karlsruhe Institute of Technology (KIT). Partnerami projektowymi natomiast są następujące instytucje:

1. Berlin School of Library and Information Science na Humboldt-Universität zu Berlin,
2. Library and Information Services department (LIS) z GFZ German Research Centre for Geosciences,
3. KIT Library z Karlsruhe Institute of Technology (KIT)
4. Biblioteka akademicka Purdue University.

W rejestrze *re3data.org* naukowcy mogą sprawdzić, jakie są warunki dostępu i korzystania z repozytorium oraz zapoznać się z opisem jego zawartości. Zgodnie z nowymi trendami w informacji wprowadzono ikony, które ułatwiają przeszukiwanie.

Rys. 2. Aspects of a Research Data Repository with the corresponding icons used in re3data.org. Autor: Heinz Pampel. Źródło: http://blogs.plos.org/tech/how-to-find-an-appropriate-research-data-repository/.

Każdy, kto chce dodać swoje repozytorium do tej listy, może skorzystać z formularza online i wypełnić *application form* przestrzegając podstawowych wymagań, które są stawiane twórcom takich archiwów czy zbiorów danych. Ważne jest, że środowiska zajmujące się danymi badawczymi zaczęły ze sobą współpracować po to, by nie powielać wysiłków. To co było do tej pory rozproszone, powoli łączy się w jeden organizm. Współpracę podjęli twórcy następujących inicjatyw: BioSharing, DataCite oraz OpenAIRE.



Rys. 3. Wzrost liczby repozytoriów surowych danych badawczych od roku 2012. [online]. [data dostępu wrzesień 2016]. Źródło danych: http://re3data.org, 2014.

7. LISTA REPOZYTORIÓW ŚWIATOWYCH A HUMANISTYKA

W rejestrze opisanym wyżej można wybrać repozytoria gromadzące dane humanistyczne, których z roku na rok jest coraz więcej. Ich liczba w podziale na dziedziny i specjalności przedstawia się następująco:

- Agricultural Economics and Sociology — 9[2]
- Ancient Cultures — 13
- Ancient History — 1
- Art History — 6
- Artificial Intelligence, Image and Language Processing — 10
- Asian Studies — 2
- Classical Archaeology — 4
- Cognitive Neuroscience and Neuroimaging — 8
- Communication Science — 8
- Early Modern History — 1
- Education Sciences — 29
- Egyptology and Ancient Near Eastern Studies — 1
- Empirical Social Research — 59
- European and American Literature — 2
- Evolution, Anthropology — 17
- Fine Arts, Music, Theatre and Media Studies — 22
- General and Applied Linguistics — 2
- General and Comparative Literature and Cultural Studies — 2
- History — 41
- Humanities — 343
- Humanities and Social Sciences — 343
- Linguistics — 43
- Literary Studies — 9
- Modern and Current History — 4
- Non-European Languages and Cultures, Social and Cultural Anthropology, Jewish Studies and Religious Studies — 14
- Political Science — 27
- Prehistory — 1
- Psychology- 13
- Religious Studies and Jewish Studies — 4
- Research on Socialization and Educational Institutions and Professions — 3
- Social Sciences — 343
- Social and Behavioural Sciences — 214
- Theology — 3
- Typology, Non-European Languages, Historical Linguistics — 2.

Wymienione repozytoria przechowują bardzo różne typy danych: od tekstów, fotografii, digitalizatów, przez ankiety, formularze do sygnałów dźwiękowych, pomiarów, grafów, wykresów czy nagrań audio/video. Ta różnorodność jest spowodowana oczywiście szerokim wachlarzem prowadzonych badań. Na stronach listy *re3data.org* wymieniono następujące typy danych: *Archived data, Audiovisual data, Configuration data, Databases, Images, Networkbased data, Plain text, Raw data, Scientific and statistical data formats, Software applications, Source code, Standard office documents, Structured graphics, Structured text, other.*

[2] Liczb podanych przy dziedzinach nie należy sumować, gdyż w analizowanym rejestrze dla każdego repozytorium można wskazać wiele dziedzin reprezentatywnych dla jego zawartości.

Dla lepszej orientacji czym zajmują się repozytoria danych badawczych w zakresie humanistyki warto przytoczyć przykładowe opisy repozytoriów humanistycznych, które pozwolą zrozumieć ich różną naturę. Opisy te są oczywiście skrócone, w rejestrze omówionym w poprzedniej sekcji można zapoznać się z ich bardziej szczegółową wersją. Poniżej zaprezentowano trzy przykłady:

Przyklad 1

Name: Bavarian Archive for Speech Signals (Bayerisches Archiv für Sprachsignale)
Subjects: Humanities Humanities and Social Sciences Linguistics.
Content types: Audiovisual data Plain text Raw data Standard office documents Structured graphics.
Countries: Germany.
Description: The Bavarian Archive for Speech Signals (BAS) is a public institution hosted by the University of Munich. This institution was founded with the aim of making corpora of current spoken German available to both the basic research and the speech technology communities via a maximally comprehensive digital speech-signal database. http://www.en.phonetik.uni-muenchen.de/research/bav_arch_spsig/index.html.

Repozytorium zawiera zarejestrowane cyfrowo wypowiedzi współczesnych Niemców wraz z opisami bibliograficznymi i treściowymi. Oto przykład do odsłuchania i przeczytania: http://www.bas.uni-muenchen.de/forschung/Bas/BasSampleseng.html.

Przyklad 2

Name: ARACHNE
Subjects: Art History Ancient Cultures Classical Archaeology Fine Arts, Music, Theatre and Media Studies History Humanities Humanities and Social Sciences
Content types: Databases Images Plain text Raw data Structured graphics other
Countries: Germany.
Description: Arachne is the central object-database of the German Archaeological Institute (DAI). In 2004 the DAI and the Research Archive for Ancient Sculpture at the University of Cologne (FA) joined the effort to support Arachne as a tool for free internet-based research. Arachne's database design uses a model that builds on one of the most basic assumptions one can make about archaeology, classical archaeology or art history: all activities in these areas can most generally be described as contextualizing objects. Arachne tries to avoid the basic mistakes of earlier databases, which limited their object modeling to specific project-oriented aspects, thus creating separated containers of only a small number of objects. All objects inside Arachne share a general part of their object model, to which a more class-specific part is added that describes the specialised properties of a category of material like architecture or topography. Seen on the level of the general part, a powerful pool of material can be used for general information retrieval, whereas on the level of categories and properties, very specific structures can be displayed. http://arachne.uni-koeln.de/drupal/?q=de.

Repozytorium zawiera fotografie obiektów archeologicznych, obiektów sztuki, mapy, opisy merytoryczne pojedynczych obiektów i całych kolekcji, np. kolekcji starożytnych zabytków Berlina.

Rys. 4. Zrzut ekranowy repozytorium Arachne. [online]. [data dostępu wrzesień 2016]. Źródło: http://arachne.uni-koeln.de/drupal/?q=de.

Przyklad 3

Name: Reading Experience Database RED.
Subjects: European and American Literature History Humanities Humanities and Social Sciences Linguistics Literary Studies
Content types: Plain text
Countries: United Kingdom
Description: RED is a collection of databases whose aim is to accumulate as much evidence as possible about reading experiences across the world. The search and browse facilities enable you to chart the reading tastes of individual readers as they travel to other countries, and consider how different environments may have affected their reading. http://www.open.ac.uk/Arts/reading/index.php.

Rys. 5. Zrzut ekranowy repozytorium Reading Experience Database RED. [online]. [data dostępu wrzesień 2016]. Źródło http://www.open.ac.uk/Arts/reading/index.php.

8. POLSKA A DANE BADAWCZE

Polska w zasadzie nie upowszechnia danych badawczych (choć są już pierwsze próby), co więcej polscy naukowcy z racji swojej nieufności do otwartości na ogół nie wyobrażają sobie dziś, że można dzielić się danymi przed publikowaniem osiągnięć na nich budowanych, co nie znaczy, że nie zmienią swojego podejścia. Instytut Biochemii i Biofizyki PAN w Warszawie ma zostać polskim ośrodkiem dla projektu ELIXIR http://www.elixir-europe.org/, w ramach którego naukowcy chcą upowszechniać dane badawcze z zakresu biologii. Przedsięwzięcie to – umieszczone na Polskiej Mapie Drogowej Infrastruktury Badawczej – jest w fazie przygotowawczej. Do chlubnych wyjątków należałoby zaliczyć następujące projekty, które zaczęły gromadzić surowe dane badawcze:

1. Narodowy Korpus Języka Polskiego http://nkjp.pl/index.php?page=15&lang=0.

Narodowy Korpus Języka Polskiego jest wspólną inicjatywą Instytutu Podstaw Informatyki PAN (koordynator), Instytutu Języka Polskiego PAN, Wydawnictwa Naukowego PWN oraz Zakładu Językoznawstwa Komputerowego i Korpusowego Uniwersytetu Łódzkiego, zrealizowaną jako projekt badawczo-rozwojowy Ministerstwa Nauki i Szkolnictwa Wyższego. Te cztery instytucje wspólnie zbudowały korpus referencyjny polszczyzny wielkości ponad półtora miliarda słów. Wyszukiwarki korpusowe (menu po prawej stronie) pozwalają przeszukiwać zasoby NKJP zaawansowanymi narzędziami uwzględniającymi odmianę polskich wyrazów, a nawet analizującymi budowę polskich zdań. […] Korpus językowy to zbiór tekstów, w którym szukamy typowych użyć słów i konstrukcji oraz innych informacji o ich znaczeniu i funkcji. Bez dostępu do korpusu nie da się dziś prowadzić badań językoznawczych, pisać słowników ani podręczników języków obcych, tworzyć wyszukiwarek uwzględniających polską odmianę, tłumaczy komputerowych ani innych programów zaawansowanej technologii językowej[3].

Korpus stanowi bazę materiałową dla nowego *Wielkiego Słownika Języka Polskiego*, tworzoną w ramach projektu badawczo-rozwojowego w Instytucie Języka Polskiego PAN. Warto zobaczyć zastosowania tego korpusu w praktyce:

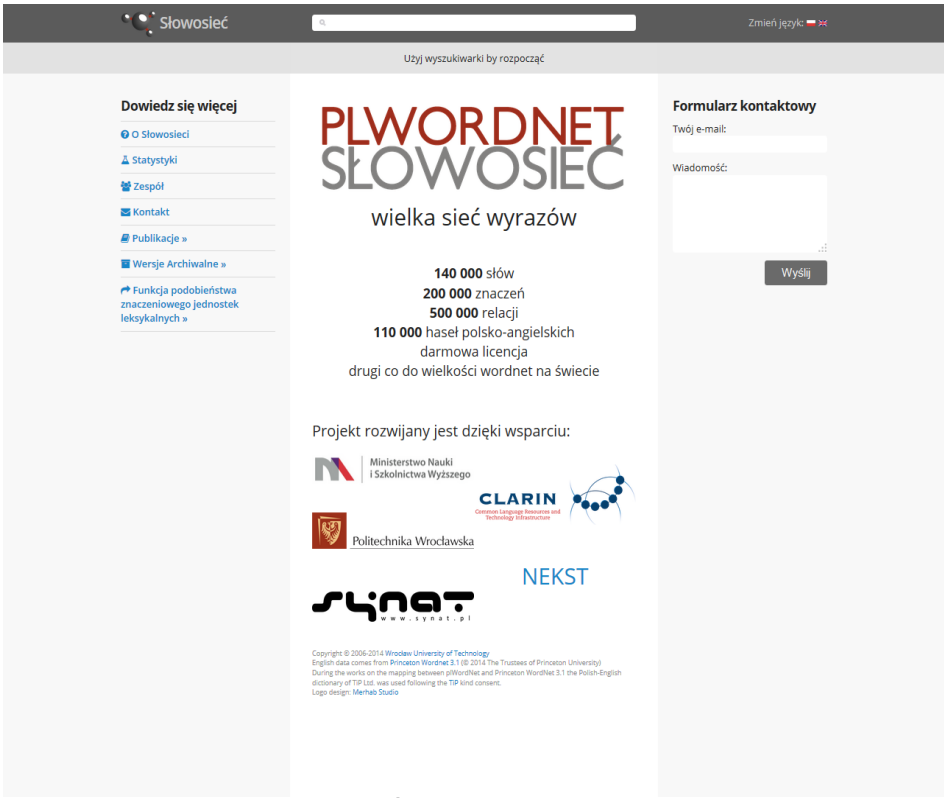
Część tekstów zebranych w ramach NKJP wykorzystywana jest na bieżąco w projekcie *Korpus Polsko-Rosyjski* afiliowanym na Wydziale Polonistyki Uniwersytetu Warszawskiego we współpracy z Uniwersytetem Pedagogicznym w Ufie oraz Narodowym Korpusem Języka Rosyjskiego. NKJP jest także wykorzystywany w wielu innych projektach realizowanych w Instytucie Podstaw Informatyki PAN oraz w jednostkach współpracujących z IPI PAN, w tym na Politechnice Wrocławskiej (m.in. przy konstrukcji kolejnych wersji Słowosieci) i w Akademii Górniczo-Hutniczej (w tym w projektach *Lingwistyczny warsztat do analizy i rozpoznawania mowy* i *System dialogowy człowiek-komputer*).[4]

Cały zasób językowy i narzędzia do wykorzystania Korpusu są dostępne na stronach internetowych projektu bez żadnych barier dostępu. Projekt ten nie przybrał formy repozytorium, jego zasób dostępny jest na stronach internetowych z różnych baz danych, ale taka jest jego natura.

2. Słowosieć — jest to interesujący projekt polegający na tworzeniu leksykalno-semantycznego korpusu języka polskiego, słownictwa powiązanego ze sobą znaczeniowo. Zasób ten jest rozwijany na Politechnice Wrocławskiej przez naukowców w tzw. Grupie Technologii Językowych G4.19. Celem projektu jest opracowanie podstawowych narzędzi i słownictwa, które będzie można wykorzystywać na wiele sposobów zarówno przez ludzi, jak i komputery. Obecnie Słowosieć zawiera ponad **140 000** słów, **200 000** znaczeń, **500 000**

[3] O projekcie NKJP. Informacja dostępna w: http://nkjp.pl/.
[4] Zastosowania Korpusu. Informacja dostępna na stronach: http://nkjp.pl/index.php?page=15&lang=0.

relacji, **110 000** haseł polsko-angielskich około 150 tys. lematów, 207 tys. jednostek leksykalnych oraz 151 tys. synsetów. W Słowosieci znajdują się rzeczowniki (116 tys.), czasowniki (17 tys.) i przymiotniki (13 tys.)[5]. Słowosieć bywa nazywana tezaurusem, słownikiem synonimów, bazą danych leksykalnych. Jest częścią większego światowego przedsięwzięcia zwanego WordNet. Dzięki połączeniu z amerykańskim Princeton WordNet Słowosieć jest również wielkim słownikiem polsko-angielskim. Słownictwo i narzędzia Słowosieci są dostępne na licencjach otwartych. Słownictwo już jest wykorzystywane w różnych projektach, np. do wspomagania pracy tłumaczy, do edukacji, do tworzenia polskiej wyszukiwarki sieciowej Nekst, rozwijanej przez Instytut Podstaw Informatyki PAN (IPI PAN) w Warszawie i Politechnikę Wrocławską.



Rys. 6. PLWordnet. Słowosieć. Źródło: http://plwordnet.pwr.wroc.pl/wordnet/.

Dane zgromadzone w Słowosieci są dostępne za darmo zarówno do celów naukowych jak i komercyjnych, mogą być wykorzystane nie tylko do budowania słowników, ale i do badań nad sztuczną inteligencją, do automatycznych tłumaczeń, do budowy polskiej wyszukiwarki itp.[6]

3. Archiwum Danych Społecznych (ADS) http://www.ads.org.pl/index.php. tworzone na Uniwersytecie Warszawskim (dane dostępne po rejestracji).

Jest to wspólne przedsięwzięcie Instytutu Studiów Społecznych UW i Instytutu Filozofii i Socjologii PAN, ale do projektu przystąpiło już pięć innych instytucji badawczych. Bardzo ważne jest to, że twórcy Archiwum Danych Społecznych (ADS) dobrze sformułowali

[5] Dane z maja 2015 roku.
[6] Werla, M.; Maryl, M. (2014) Humanistyczne projekty cyfrowe w Polsce. Warszawa Poznań [dostęp: 20.04.15]. Dostępny w WWW: http://lib.psnc.pl/Content/655/Humanistyczne_projekty_cyfrowe_w_Polsce_final.pdf.

korzyści płynące z gromadzenia zbiorów danych badawczych W swoim podręczniku ujmują je tak:

1. ułatwienie realizacji postulatu otwartości warsztatu badawczego i tym samym wdrożenie idei intersubiektywnej kontroli procesu badawczego.
2. inspirowanie do mnożenia analiz i hipotez z wykorzystaniem zebranych i dostępnych już danych.
3. promowanie nowych badań i umożliwienie testowania nowych lub alternatywnych metod weryfikacji postawionych już bądź stawianych hipotez.
4. usprawnianie metod zbierania danych i konstrukcji pomiarów. Ogólnodostępne archiwum danych otwiera naukowej społeczności możliwość wypracowywania standardów metodologicznych
5. dostęp do danych z badań już zrealizowanych może się przyczynić do poszerzenia zakresu dokonywanych analiz bez konieczność powtórnego przeprowadzania badań. Nie istniałaby ogromna liczba publikacji i artykułów powstałych na bazie takich badań jak np. Polskie Generalne Sondaże Społeczne, gdyby ich autorzy musieli sami zbierać tego typu dane.
6. archiwum jest wreszcie nieporównywalnym z żadnym innym źródłem danych wykorzystywanych dla celów dydaktycznych. Dostarcza zarówno wykładowcom jak i studentom dane o najwyższym standardzie metodologicznym. (Fragment Podręcznika archiwizacji danych społecznych, dostępny pod adresem: http://www.ads.org.pl/index.php?tresc=podrecznikADS.html.[7]

Rys. 7. Zrzut ekranowy repozytorium ADS. Źródło: http://www.ads.org.pl/.

Niestety, nie wszystkie dane badawcze są w tym repozytorium otwarte, ponieważ nie wszystkie instytucje współpracujące sobie tego życzą. Zasady ustala się na podstawie odrębnych umów. Pobieranie zgromadzonych w ADS danych przez indywidualnego użytkownika wymaga uprzedniej rejestracji, co też stanowi utrudnienie. Ale po takiej rejestracji można nieodpłatne pobierać otwarte dane. Do zastrzeżonych zasobów dostęp mają jedynie instytucje, które zawarły uprzednio z Archiwum Umowę o współpracy i odprowadziły opłatę.

[7] F, B.; Jerzyński, T.; Zieliński, M. (2004). Podręcznik archiwizacji danych społecznych. Zespół Ośrodka Badań Socjologicznych, Instytutu Studiów Społecznych, Uniwersytetu Warszawskiego [dostęp: 20.04.15]. Dostępny w WWW: http://www.ads.org.pl/pdf/podrecznikADS.pdf.

4. Archiwum Historii Mówionej Ośrodka Karta to zbiór relacji biograficznych (około 5 tys. nagrań audio i 120 wideo) oraz innych archiwalnych świadectw XX wieku. Pierwsze nagrania pochodzą z 1987 r. i oczywiście Ośrodek nada kontynuuje zbieranie innych (http://www.audiohistoria.pl/).

Rys. 8. Zrzut ekranowy Archiwum Historii mówionej. Źródło: http://www.audiohistoria.pl/.

Głównym celem — jak piszą na swoich stronach internetowych twórcy archiwum —

jest utrwalenie pamięci odchodzących generacji. W tym celu angażujemy do nagrywania i opracowywania relacji kolejne środowiska i współpracowników z całej Polski. Nasza praca w dużym stopniu możliwa jest także dzięki wsparciu finansowemu wielu instytucji, w tym Ministerstwa Kultury i Dziedzictwa Narodowego, Ministerstwa Nauki i Szkolnictwa Wyższego, Komisji Europejskiej oraz Kancelarii Senatu RP.

Trzeba przyznać, że Ośrodek wykonał bardzo dużą pracę i nie jest to jego jedyne archiwum surowych danych. Twórcy gromadzą także fotografie i dokumenty w wielu bazach danych, które są dostępne online na stronie http://www.karta.org.pl/Archiwa_i_bazy_danych/78, bez żadnych barier - co jest niezwykle cenne i jak wcześniej pisałam w Polsce niekoniecznie popularne. Postulat, żeby wszystkie zasoby wytwarzane za publiczne pieniądze były dla obywatela otwarte bez barier technicznych, ekonomicznych czy prawnych jest z trudem realizowany w polskiej rzeczywistości instytucjonalnej.

W 2014 r. w Ministerstwie Nauki i Szkolnictwa Wyższego pod egidą ministra profesora Włodzisława Ducha rozpoczęły się debaty na temat wdrożenia w Polsce modeli open access dla nauki polskiej w tym danych badawczych. 20 marca 2015 roku MNiSW powołało Zespół doradczy do spraw otwartego dostępu do treści naukowych (http://www.nauka.gov.pl/g2/oryginal/2015_05/f0061d2ae21e462a5816f8a8cbe4fdfb.pdf), który opracował politykę open access dla Polski. Oficjalny dokument nazywa się „Kierunki rozwoju otwartego dostępu do treści naukowych w Polsce" i był opublikowany w październiku 2015 (http://www.nauka.gov.pl/komunikaty/kierunki-rozwoju-otwartego-dostepu-do-publikacji-i-wynikow-badan-naukowych-w-polsce.html). Problem otwartych

danych badawczych jest w nim tylko zasygnalizowany jako ważny i taki, którym należy się zająć w najbliższych latach.

9. PODSUMOWANIE I WNIOSKI

Podsumowując przegląd repozytoriów danych badawczych dla humanistyki należy podkreślić, że wiele krajów europejskich chętnie dzieli się swoimi danymi tak, by inne ośrodki mogły z nich korzystać. Niestety, nie możemy tego powiedzieć o Polsce. W Polsce zarówno wprowadzanie zasad powszechnego udostępniania publikacji naukowych, jak i danych badawczych nie przebiega łatwo. Mamy bardzo wiele przykładów otwierania repozytoriów publikacji naukowych (ponad 20) i budowania platform czasopism otwartych (kilka), natomiast repozytoria danych badawczych można policzyć na palcach jednej ręki, choć - jak wynika z przeglądów naukowych baz danych prezentowanych na konferencjach naukowych takich jak INFOBAZY - nie jest ich mało, ale są one ukryte na serwerach poszczególnych uczelni czy instytutów krajowych. Są nie tylko ukryte, ale często zamknięte dla powtórnego ich użycia i to z mocy ustaw. Oto przykład Państwowego Instytutu Geologicznego i jego **Centralnej Bazy Danych Geologicznych:**

Wgląd i udostępnienie informacji geologicznej od 1 stycznia 2014 r. realizowany jest w oparciu o nowe prawo geologiczne i górnicze oraz rozporządzenia. Rozporządzenie określa wgląd jako nieodpłatne zapoznanie się ze zgromadzoną informacją geologiczną, bez prawa dokonywania reprodukcji, odpisu, odrysu, wydruku, fotokopii lub kopii w postaci elektronicznej dokumentów i zbiorów danych, a także bez prawa pobierania próbek (§ 9.1 — rozporządzenie Ministra Środowiska z dnia 15 grudnia 2011 r. w sprawie gromadzenia i udostępniania informacji geologicznej)[http://geoportal.pgi.gov.pl/cbdg/dane/dostep].

Tworzenie tych danych jest z pewnością finansowane z pieniędzy podatników, ale podatnik nie może liczyć na pełne ich wykorzystanie, tym bardziej komercyjny, który płaci jeszcze większe podatki. Jaki zatem mamy zwrot z inwestycji, którą poczyniło nasze państwo? Żaden. Dlaczego w Polsce ukrywa się tak ważne dane? Z powodów strategicznych, bezpieczeństwa, z chęci ponownego zysku? Czy nie większym zyskiem byłoby, gdyby firmy i obywatele mogli takie dane wykorzystać i wprowadzić nowe usługi, przedsięwziąć interesujące inwestycje, rozwijać kraj? Pytania te i podobne stawia sobie cały świat. Warto na nie odpowiedzieć także i w Polsce, szczególnie dziś, gdy Komisja Europejska wprowadza konkretne inicjatywy i zaleca coraz szerszą transparentność w obszarze administracji i nauki. W programie Horyzont 2020 wprowadzono test dla gromadzenia i upowszechniania danych badawczych, co oznacza, że jeśli się uda, za parę lat będziemy wszyscy zobligowani do podobnych działań.

BIBLIOGRAFIA

1. Bednarek-Michalska, B. (2012). Repozytoria surowych danych — dlaczego biblioteki powinny je znać? *Biuletyn EBIB* [online] nr 8 (135) [dostęp: 20.04.15]. Dostępny w WWW: http://www.nowyebib.info/images/stories/numery/135/135_michalska_.pdf.
2. Bednarek-Michalska, B. (2014). Dlaczego open access w Horyzoncie 2020? [online]Uwolnijnaukę.pl [wpis z dania 1 sierpnia 2014], [dostęp:20.04.15]. Dostępny w WWW: http://uwolnijnauke.pl/open-access-w-horyzoncie-2020/.
3. F, B.; Jerzyński, T.; Zieliński, M. (2004). Podręcznik archiwizacji danych społecznych. Zespół Ośrodka Badań Socjologicznych, Instytutu Studiów Społecznych, Uniwersytetu Warszawskiego [dostęp: 20.04.15]. Dostępny w WWW: http://www.ads.org.pl/pdf/podrecznikADS.pdf.

4. EC (2013b). Commission launches pilot to open up publicly funded research data (2013, 16 grudzień). European Commission. Press Release Database [dostęp: 20.04.15]. Dostępny w WWW: http://europa.eu/rapid/press-release_IP-13-1257_en.htm.
5. EC (2013b). Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020 v.1.0 [online]. European Commission. Research and Innovation [dostęp: 20.04.15]. Dostępny w WWW: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf.
6. Euroalert (2011). Commission will adopt measures for an open data strategy [online]. Current news from European Union. Euroalert.net [dostęp: 20.04.15]. Dostępny w WWW: http://euroalert.net/en/news.aspx?idn=13979.
7. G8 (2013). G8 Science Ministers Statement, on 12 June the Royal Society hosted the first ever G8 joint Science Ministers and national science academies meeting in London [online]. Foreign & Commonwealth Office, 13.07.13 [dostęp: 20.04.15]. Dostępny w WWW: https://www.gov.uk/government/news/g8-science-ministers-statement.
8. Murray-Rust, P.; Neylon, C.; Pollock, R.; Wilbanks, J. (2010). Panton Principles, Principles for open data in science [online] Paton Principles. [data dostępu 20.04.15]. Dostępny w: http://pantonprinciples.org.
9. McDonald, D.; Kelly, U. (2012) Viewforth Consulting Value and benefits of text mining [online]. Report. JISC England 2012 [dostęp: 20.04.15]. Dostępny w WWW: http://www.webarchive.org.uk/wayback/archive/20140613225457/http://www.jisc.ac.uk/reports/value-and-benefits-of-text-mining.
10. Pampel, H. (2013). How to find an appropriate research data reposit[online]. Posted on blog: November 4 [dostęp: 20.04.15]. Dostępny w WWW: http://blogs.plos.org/tech/how-to-find-an-appropriate-research-data-repository/.
11. Pampel, H.; Vierkant P., Scholze F., Bertelmann R., Kindling M., et al. (2013). Making Research Data Repositories Visible: The re3data.org Registry[online]. PLoS ONE 8(11): e78080. doi:10.1371/journal.pone.0078080 [dostęp: 20.04.15]. Dostępny w WWW: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0078080.
12. Poynder, R. (2012). A New Declaration of Rights: Open Content Mining [online]. Wywiad i wpis w blogu [data dostępu 20.04.15]. Dostępny w: http://www.richardpoynder.co.uk/Content_Mining.pdf.
13. re3data.org (2014). Over 1,000 research data repositories indexed in re3data.org [online]. Aktualności z dnia 20.11.14. Wpis z portalu [dostęp:20.04.15]. Dostępny w WWW: http://www.re3data.org/2014/11/over-1000-research-data-repositories-indexed-in-re3data-org/
14. Stebbins, Michael (2013). *Expanding Public Access to the Results of Federally Funded Research. Washington* 22.02.13 [online]. Office of Science and Technology Policy [dostęp: 20.04.15]. Dostępny w WWW: https://www.whitehouse.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research.
15. Tenopir, C.; Allard, S.; Douglass, K.; Aydinoglu, A.U.; Wu, L.; Read, E.; Manoff, M.; Frame, M. (2011). Data Sharing by Scientists: Practices and Perceptions [online]. PLoS One DOI: 10.1371/journal.pone.0021101. [data dostępu 20.04.15]. Dostępny w: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0021101.
16. Werla, M.; Maryl, M. (2014) Humanistyczne projekty cyfrowe w Polsce. Warszawa Poznań [dostęp: 20.04.15]. Dostępny w WWW: http://lib.psnc.pl/Content/655/Humanistyczne_projekty_cyfrowe_w_Polsce_final.pdf.

Artykuł jest sponsorowany przez Narodowe Centrum Nauki przez grant o numerze 2013/11/B/HS2/03048/ zatytułowany: „Badanie struktury i dynamiki rozwoju cyfrowych zasobów wiedzy przy pomocy metod wizualizacji”.