ZAGADNIENIA
NAUKOZNAWSTWA
4 (206), 2015
PL ISSN 0044 – 1619

Adam Jachimczyk
Instytut Dziennikarstwa i Informacji
Uniwersytet Jana Kochanowskiego w Kielcach

Otwarte dane badawcze. Casus polskich instytutów badawczych

Abstrakt. Przedmiotem artykułu są zasoby danych badawczych tworzonych i udostępnianych w Internecie przez polskie instytuty naukowo-badawcze. Badanie miało dwa cele: 1) określenie liczby zasobów tworzonych przez instytuty badawcze oraz rodzaju gromadzonych w nich danych; 2) zbadanie stopnia dostępności analizowanych zasobów dla innych użytkowników.

Analiza wykazała dość wysoki odsetek instytutów, które nie tworzą żadnych zasobów danych badawczych. Ponadto z badania wynika, że wśród rodzajów udostępnianych zasobów przeważają zbiory metadanych. Barierą blokującą dostęp i wykorzystanie udostępnionych zasobów jest brak informacji na temat warunków ponownego użycia danych oraz ograniczanie ich wykorzystania przez innych użytkowników tylko do własnego użytku.

Słowa kluczowe: instytuty naukowo-badawcze, Internet, otwarte dane badawcze

Open research data. The Polish scientific-research institutes' casus

Abstract. The subject of the article is research data resources created and disseminated on the Internet by Polish scientific-research institutes. The aim of the research was: 1) identification of the number of the resources and the type of the gathered data; 2) analysis of the degree of accessibility of the resources to different users.

The research demonstrated quite a high ratio of the institutes which do not create any resources. Furthermore, within the available data, metadata sets are the most dominant. The lack of information about the rules of data re-use and limiting it only for fair use by other users are the barriers blocking the re-use of the research data available.

Keywords: scientific-research institutes, Internet, open research data

Wstęp

Współczesną naukę zaczyna charakteryzować coraz większa otwartość, której najbardziej widocznym przejawem jest rozwój ruchu open access promującego swobodny dostęp do publikacji naukowych w Internecie. W ramach tego ruchu podejmuje się także działania na rzecz bezpłatnego udostępniania w Sieci danych badawczych (research data) gromadzonych w ramach badań finansowanych ze środków publicznych (Murray-Rust 2008, 55).

Dane badawcze obejmują każdy materiał wykorzystywany w badaniach naukowych. Zalicza się do nich nie tylko dane generowane przez placówki naukowo-badawcze, ale także dane wykorzystywane do celów badawczych wytwarzane przez lub dla jednostek administracji publicznej, pierwotnie do celów nie związanych z nauką (Uhlir, Schröder 2007, 36; Wessels i in. 2014, 51). Wprawdzie określe-

nie „dane" sugeruje, że chodzi tylko o nieprzetworzone, tzw. surowe dane[1], ale w praktyce do kategorii danych badawczych często zalicza się również publikacje, materiały dźwiękowe, graficzne, filmowe, czy różnego typu artefakty (Wessels i in. 2014, 51).

Wprawdzie już ponad 10 lat temu sygnatariusze Deklaracji Berlińskiej, a nieco później *Organisation for Economic Co-operation and Development* (OECD) i Komisja Europejska poparły ideę otwartego dostępu do danych badawczych, zwłaszcza tych, które zostały zebrane w czasie badań finansowanych ze środków publicznych, ale w porównaniu z dużą grupą ogólnie dostępnych repozytoriów publikacji naukowych, liczba baz danych gromadzących i udostępniających dane badawcze jest zauważalnie mniejsza (Deklaracja 2005; EU 2011; OECD 2007).

Nie ma jednej przyczyny, która hamuje dostęp do danych badawczych innym naukowcom. W literaturze zwraca się uwagę np. na specyfikę funkcjonowania współczesnego systemu wydawniczego, który koncentruje się bardziej na wydawaniu tekstów naukowych, a nieco zaniedbuje możliwość przechowywania i udostępniania danych badawczych. Ponadto wydawcy komercyjni nakładają na rozpowszechnianie danych te same ograniczenia, które towarzyszą udostępnianiu publikacji naukowych (Molloy 2011, 1).

Część środowiska naukowego niechętnie odnosi się do dzielenia się danymi z innymi naukowcami. Opracowanie danych badawczych jest bowiem czasochłonne a ich udostępnienie nie jest traktowane jako publikacja naukowa i w ogólnym rozrachunku nie decyduje o kolejnych stopniach awansu naukowego (Borgman 2010, 7). Dzieleniu się danymi towarzyszy także obawa, że inni badacze efektywniej je wykorzystają (Reichman i in. 2011, 704).

Przeszkodą jest również brak narodowych uregulowań prawnych, które precyzowałyby zasady dostępu i wykorzystania danych badawczych gromadzonych w ramach projektów finansowanych ze środków publicznych (EU 2011, 23). Ponadto naukowcy wskazują na problemy związane z zapewnieniem finansowania dla projektów informatycznych mających na celu udostępnienie danych w Internecie (EU 2012, 5).

Mimo wskazanych problemów postulat otwartego dostępu do danych badawczych wyrażany jest w coraz większej liczbie publikacji akcentujących korzyści płynące z dzielenia się danymi. Dla zwolenników otwarcia danych istotne jest m.in. obniżenie kosztów pracy naukowej, gdyż dostępność danych w Internecie redukuje konieczność ponownego gromadzenia materiału badawczego (Hofmokl i in. 2009, 59). Otwarcie danych wpływa również na poprawę jakości prac naukowych. Pozwala bowiem innym naukowcom prowadzić badania w oparciu o ten

[1] W piśmiennictwie naukowym rozróżnia się terminy dane i informacja. Dane to obiektywne fakty (np. statystyki), a informacja to dane, ale ustrukturyzowane i zinterpretowane. Zob. m.in. K. Materska, 2007, *Informacja w organizacjach społeczeństwa wiedzy*, Warszawa: Stowarzyszenie Bibliotekarzy Polskich, s. 47; W. Babik, 2008, *Informacja naukowa jako przedmiot zarządzania*, w: Zarządzanie informacją w nauce, pod redakcją Diany Pietruch-Reizes, Katowice: Wydawnictwo Uniwersytetu Śląskiego, s. 43.

sam materiał, co sprzyja wychwytywaniu nierzetelności w nauce i wykrywaniu przypadków fabrykowania danych badawczych (Tenopir i in. 2011, 1). Dla środowiska ważna jest również możliwość nawiązania współpracy naukowej między badaczami, testowanie nowych hipotez i koncepcji badawczych, jak również okazja do tworzenia nowych zasobów, powstających z łączenia danych pochodzących z różnych źródeł (Uhlir, Schröder 2007, 43).

Przedmiot badania

Dyskusja na temat danych badawczych skłania do podjęcia analiz dotyczących udostępniania ich w Internecie przez polskie placówki naukowe. Kwestia ta nie znalazła bowiem szerszego odzwierciedlenia w piśmiennictwie naukowym. Remigiusz Sapa, analizując uczelniane repozytoria zarejestrowane w Federacji Bibliotek Cyfrowych, zwrócił uwagę, że żadne z nich nie oferowało dostępu do zbiorów danych pochodzących z badań (Sapa 2013, 126). Z kolei autorzy opracowania *„Otwarta nauka w Polsce 2014"* podkreślili przeszkody towarzyszące tworzeniu repozytoriów danych badawczych pisząc, że ruch na rzecz otwarcia danych „… napotyka ponadto problemy praktycznie nie występujące w odniesieniu do publikacji, na przykład dotyczące dostępu i ponownego wykorzystania informacji publicznej czy wynikające z konieczności poszanowania prywatności i zasad ochrony danych osobowych" (Leśniak i in., 2014, 26).

W artykule przedmiotem analizy stały się zasoby danych tworzonych i udostępnianych na WWW przez polskie instytuty naukowo-badawcze. W Polsce działa obecnie 116 instytutów badawczych, które zatrudniają ok. 40 tys. pracowników (ORGIB 2014).

Ich status określa ustawa z dnia 30 kwietnia o instytutach badawczych (Ustawa 2010). Instytutem badawczym jest wyodrębniona pod względem prawnym, organizacyjnym i ekonomiczno-finansowym jednostka prowadząca badania naukowe i prace rozwojowe ukierunkowane na ich wdrożenie i zastosowanie w praktyce. Poza prowadzeniem badań naukowych, instytuty mogą m.in. upowszechniać wyniki badań naukowych i prac rozwojowych; prowadzić i rozwijać bazy danych związane z przedmiotem działania instytutu; prowadzić działalność w zakresie informacji naukowej, technicznej i ekonomicznej, wynalazczości oraz ochrony własności przemysłowej i intelektualnej, a także wspierającej innowacyjność przedsiębiorstw; prowadzić działalność wydawniczą związaną z prowadzonymi badaniami naukowymi i pracami rozwojowymi. Ustawa nie nakłada jednak obowiązku udostępniania tworzonych baz danych.

Metodologia

Badanie, przeprowadzone między majem a lipcem 2015 r., polegało na analizie stron internetowych instytutów badawczych i zidentyfikowaniu informacji o two-

rzonych i udostępnianych zasobach danych (gromadzonych w bazach danych, jak również dostępnych na stronach WWW w postaci plików w formacie pdf czy xls). Kategorię danych badawczych potraktowano bardzo szeroko obejmując nią surowe dane, metadane, pełne teksty dokumentów, materiały graficzne i video[2].

Celem analizy było:

- określenie liczby zasobów tworzonych i udostępnianych przez wspomniane jednostki,
- określenie rodzaju danych gromadzonych w zasobach (surowe dane badawcze, metadane, pełne teksty, materiały graficzne i video),
- zbadanie stopnia otwarcia zasobów dla innych użytkowników. Kwestię tę, kierując się kryteriami otwartej wiedzy i tzw. Zasadami Pantone (DOW 2015; Murray-Rust i in. 2010), zbadano analizując trzy zagadnienia:
  a. dostępność w Internecie danych badawczych. Idea otwartych danych zakłada, że znajdują się one w domenie publicznej i są bezpłatnie dostępne dla ogółu użytkowników (Murray-Rust i in. 2010)[3]. W badaniu sprawdzono dostępność w Sieci zasobów tworzonych przez instytuty badawcze oraz ograniczenia w dostępie do nich. Zbiory danych podzielono na trzy grupy: dostępne bez ograniczeń, dostępne z ograniczeniami, niedostępne w Internecie. Zdefiniowano następujące rodzaje ograniczeń: płatny dostęp, konieczność rejestracji w celu uzyskania dostępu do zasobu, dostęp tylko dla pracowników danego instytutu, dostęp tylko dla uprawnionych osób;
  b. format udostępnianych danych badawczych. Przyjmuje się, że otwarte dane to takie, które są dostępne w otwartym formacie pozwalającym na ich przetwarzanie za pomocą ogólnie dostępnego oprogramowania typu open source (DOW 2015). Można do tego kryterium dodać też łatwość przetwarzania dostępnych danych (Molloy 2011, 1). Na przykład, znacznym utrudnieniem dla użytkownika będzie przetwarzanie w arkuszu kalkulacyjnym danych numerycznych znajdujących się w pliku w formacie html czy pdf, gdyż zmusza go to do zastosowania dodatkowego oprogramowania konwertującego dane z tych formatów do formatu akceptowanego przez arkusz kalkulacyjny. Badanie pozwoliło na zidentyfikowanie formatów, w jakich dane zostały udostępnione użytkownikom. Zbadano również, czy mogą je przetwarzać przy pomocy darmowego dostępnego w Internecie oprogramowania;
  c. warunki wykorzystania zasobów. Otwarty dostęp zakłada, że licencja nie powinna narzucać żadnych ograniczeń natury prawnej lub technicznej

[2] Podczas badania nie rozstrzygano kwestii, czy odszukany zbiór danych zostanie wykorzystany w pracy naukowej. Każdy zasób został potraktowany jako potencjalny materiał badawczy, np. zbiory danych bibliograficznych mogą posłużyć do badań o charakterze bibliometrycznym.

[3] Dopuszcza się pewną odpłatność, ale cena nie może być wyższa niż koszt jednorazowej reprodukcji, por. (DOW 2015).

co do sposobu i celu ich wykorzystania (DOW 2015; Murray-Rust i in. 2010). W zasadzie jedynym warunkiem powinien być obowiązek oznaczenia autora zbioru danych. Zbadano dostępność informacji na temat warunków korzystania ze zbiorów danych oraz przenalizowano je pod kątem ograniczeń nakładanych na użytkowników w zakresie pobierania, przetwarzania i rozpowszechniania danych.

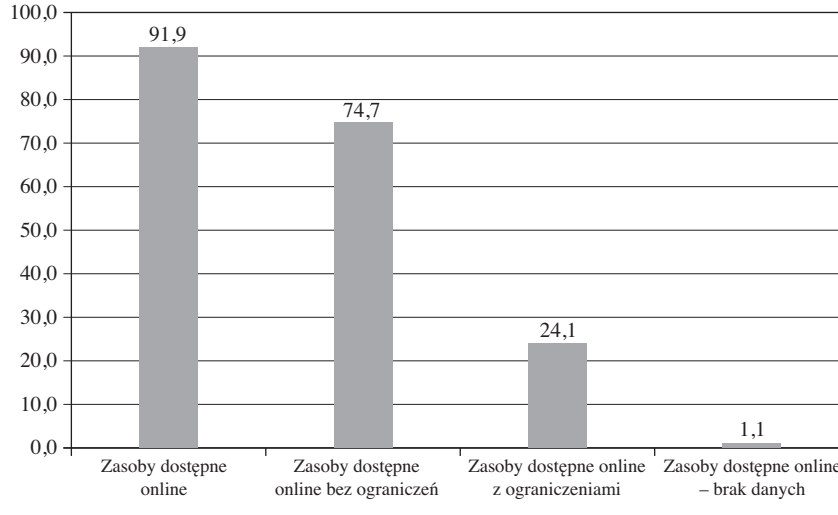
W analizie nie uwzględniono pojedynczych wydawnictw periodycznych[4] i książek[5] publikowanych w formie elektronicznej i udostępnianych w Internecie przez instytuty badawcze.

Wyniki analizy

Zasoby danych

Ogółem zasoby różnego rodzaju danych tworzą 82 (70,7%) instytuty badawcze. Podczas analizy, jak ilustruje wykres 1, zidentyfikowano 383 zbiory danych, z których prawie 92% jest dostępne w Internecie. Nieco mniejszy odsetek – niecałe 75% – przypada na zasoby, które są dostępne bez żadnych ograniczeń w Sieci.

Wykres 1. Odsetek zasobów danych udostępnianych online przez instytuty badawcze



[4] Przeprowadzona na potrzeby innego badania analiza stron internetowych wykazała, że ogółem prawie 65% wydawnictw periodycznych wydawanych przez instytuty badawcze udostępnia w Internecie swoje bieżące wydania. Prawie 94% wydawnictw publikuje swoją zawartość w formacie pdf. Ponad połowa wydawnictw nie informuje o warunkach korzystania z czasopisma. 25 czasopism udostępnia treść na licencji Creative Commons, 39 czasopism stosuje standardową formułę zastrzeżenia praw autorskich copyright.

[5] 22 instytuty badawcze udostępniają, na swojej stronie WWW lub w serwisie *e-Publikacje Nauki Polskiej* (https://www.epnp.pl/), pełne teksty książek wyłącznie w formacie pdf. Nie jest to jednak pełna oferta wydawnicza.

Wśród instytutów badawczych, jak pokazuje tab. 1, najwięcej zasobów tworzy i udostępnia Państwowy Instytut Geologiczny (PIG) oraz Ośrodek Przetwarzania Informacji (OPI). Z kolei Instytut Badawczy Leśnictwa tworzy ponad 50 baz danych, ale dostęp do przeważającej części z nich zastrzega tylko dla własnych pracowników. Łącznie te trzy instytuty tworzą 33,4% wszystkich zidentyfikowanych zasobów. W przypadku zbiorów dostępnych online bez ograniczeń ten odsetek jest już mniejszy i wynosi 21% ogółu znalezionych zasobów. Ponad połowa instytutów (51,2%) tworzy tylko jeden zbiór danych. W tej grupie odsetek jednostek udostępniających zasoby bez ograniczeń w Internecie wynosi 81%.

Tabela 1. Liczba zasobów tworzonych i udostępnianych przez instytuty naukowo-badawcze

Lp.	Nazwa jednostki	Ogółem	Dostępne online	Dostępne online bez ograniczeń
1.	Instytut Badawczy Leśnictwa	57	57	5
2.	Państwowy Instytut Geologiczny. Państwowy Instytut Badawczy	38	38	37
3.	Ośrodek Przetwarzania Informacji. Państwowy Instytut Badawczy	33	33	32
4.	Główny Instytut Górnictwa w Katowicach	22	17	1
5.	Instytut Geodezji i Kartografii w Warszawie	17	6	5
6.	Instytut Logistyki i Magazynowania w Poznaniu	13	13	13
7.	Centralny Instytut Ochrony Pracy. Państwowy Instytut Badawczy	12	12	12
8.	Instytut Meteorologii i Gospodarki Wodnej. Państwowy Instytut Badawczy	12	12	11
9.	Instytut Zaawansowanych Technologii Wytwarzania w Krakowie	12	12	12
10.	Instytut Zootechniki. Państwowy Instytut Badawczy w Krakowie	11	11	11
11.	Instytut Technologii Drewna w Poznaniu	10	3	3
12.	Instytut Chemicznej Przeróbki Węgla w Zabrzu	9	9	6
13.	Instytut Techniki Budowlanej w Warszawie	9	9	9
14.	Instytut Medycyny Pracy im. prof. dra med. Jerzego Nofera w Łodzi	7	7	6
15.	Instytut „Pomnik – Centrum Zdrowia Dziecka” w Warszawie	5	5	5
16.	Instytut Ochrony Roślin. Państwowy Instytut Badawczy w Poznaniu	5	5	4
17.	Narodowy Instytut Zdrowia Publicznego – Państwowy Zakład Higieny w Warszawie	5	5	5
18.	Instytut Badań Edukacyjnych	4	4	4

Lp.	Nazwa jednostki	Ogółem	Dostępne online	Dostępne online bez ograniczeń
19.	Instytut Biotechnologii Przemysłu Rolno-Spożywczego im. prof. Wacława Dąbrowskiego w Warszawie	4	3	3
20.	Instytut Ciężkiej Syntezy Organicznej „Blachownia" w Kędzierzynie-Koźlu	4	4	4
21.	Instytut Mechanizacji Budownictwa i Górnictwa Skalnego	4	4	4
22.	Instytut Ekologii Terenów Uprzemysłowionych	3	3	2
23.	Instytut Ekonomiki Rolnictwa i Gospodarki Żywnościowej. Państwowy Instytut Badawczy w Warszawie	3	3	3
24.	Instytut Hodowli i Aklimatyzacji Roślin. Państwowy Instytut Badawczy w Radzikowie	3	3	3
25.	Instytut Łączności. Państwowy Instytut Badawczy w Warszawie	3	3	3
26.	Instytut Odlewnictwa w Krakowie	3	3	1
27.	Instytut Psychiatrii i Neurologii w Warszawie	3	3	3
28.	Instytut Spawalnictwa w Gliwicach	3	2	1
29.	Instytut Technologii Materiałów Elektronicznych w Warszawie	3	3	3
30.	Instytut Włókiennictwa w Łodzi	3	3	3
31.	Wojskowy Instytut Medyczny	3	3	3
32.	Instytut Energetyki w Warszawie	2	1	1
33.	Instytut Fizjologii i Patologii Słuchu w Warszawie	2	2	2
34.	Instytut Gruźlicy i Chorób Płuc w Warszawie	2	2	2
35.	Instytut Matki i Dziecka w Warszawie	2	2	2
36.	Instytut Mechaniki Precyzyjnej w Warszawie	2	0	0
37.	Instytut Metali Nieżelaznych w Gliwicach	2	2	2
38.	Instytut Nafty i Gazu. Państwowy Instytut Badawczy w Krakowie	2	2	1
39.	Instytut Nowych Syntez Chemicznych w Puławach	2	2	1
40.	Instytut Ochrony Środowiska	2	1	1
41.	Centralne Laboratorium Ochrony Radiologicznej w Warszawie	1	1	1
42.	Centrum Naukowo-Badawcze Ochrony Przeciwpożarowej im. Józefa Tuliszkowskiego. Państwowy Instytut Badawczy	1	1	1
43.	Centrum Onkologii – Instytut im. Marii Skłodowskiej-Curie w Warszawie	1	1	1

Tabela 1 – c.d.

Lp.	Nazwa jednostki	Ogółem	Dostępne online	Dostępne online bez ograniczeń
44.	Instytut „Centrum Zdrowia Matki Polki” w Łodzi	1	1	1
45.	Instytut Badań Rynku, Konsumpcji i Koniunktur w Warszawie	1	1	1
46.	Instytut Badawczy Dróg i Mostów w Warszawie	1	1	1
47.	Instytut Biopolimerów i Włókien Chemicznych w Łodzi	1	1	1
48.	Instytut Ceramiki i Materiałów Budowlanych w Warszawie	1	1	1
49.	Instytut Chemii Przemysłowej imienia Profesora Ignacego Mościckiego w Warszawie	1	1	1
50.	Instytut Europy Środkowo-Wschodniej	1	1	1
51.	Instytut Fizyki Plazmy i Laserowej Mikrosyntezy im. Sylwestra Kaliskiego w Warszawie	1	1	1
52.	Instytut Hematologii i Transfuzjologii w Warszawie	1	1	1
53.	Instytut Kardiologii im. Prymasa Tysiąclecia Stefana Kardynała Wyszyńskiego w Warszawie	1	1	1
54.	Instytut Lotnictwa w Warszawie	1	1	1
55.	Instytut Maszyn Matematycznych w Warszawie	1	1	1
56.	Instytut Metalurgii Żelaza im. Stanisława Staszica	1	1	1
57.	Instytut Morski w Gdańsku	1	1	0
58.	Instytut Napędów i Maszyn Elektrycznych KOMEL w Katowicach	1	1	1
59.	Instytut Obróbki Plastycznej w Poznaniu	1	1	1
60.	Instytut Ogrodnictwa w Skierniewicach	1	1	1
61.	Instytut Optyki Stosowanej imienia prof. Maksymiliana Pluty w Warszawie	1	1	1
62.	Instytut Reumatologii im. prof. dr hab. med. Eleonory Reicher w Warszawie	1	1	1
63.	Instytut Sportu w Warszawie	1	1	1
64.	Instytut Techniki Górniczej KOMAG w Gliwicach	1	0	0
65.	Instytut Techniki i Aparatury Medycznej ITAM w Zabrzu	1	1	1
66.	Instytut Technologiczno-Przyrodniczy w Falentach	1	1	1
67.	Instytut Tele- i Radiotechniczny w Warszawie	1	1	1
68.	Instytut Transportu Samochodowego w Warszawie	1	1	0
69.	Instytut Warzywnictwa w Skierniewicach	1	1	1

Lp.	Nazwa jednostki	Ogółem	Dostępne online	Dostępne online bez ograniczeń
70.	Instytut Żywności i Żywienia im. prof. dra med. Aleksandra Szczygła w Warszawie	1	1	0
71.	Morski Instytut Rybacki. Państwowy Instytut Badawczy	1	0	0
72.	Narodowe Centrum Badań Jądrowych w Otwocku-Świerku	1	1	0
73.	Narodowy Instytut Leków w Warszawie	1	1	1
74.	Naukowa i Akademicka Sieć Komputerowa w Warszawie	1	1	1
75.	Ośrodek Badań Naukowych im. W. Kętrzyńskiego w Olsztynie	1	1	1
76.	Państwowy Instytut Naukowy – Instytut Śląski w Opolu	1	1	1
77.	Poltegor-Instytut. Instytut Górnictwa Odkrywkowego we Wrocławiu	1	1	0
78.	Przemysłowy Instytut Maszyn Rolniczych w Poznaniu	1	1	0
79.	Wojskowy Instytut Chemii i Radiometrii w Warszawie.	1	1	1
80.	Wojskowy Instytut Łączności im. prof. dr. hab. Janusza Groszkowskiego w Zegrzu	1	1	1
81.	Wojskowy Instytut Medycyny Lotniczej	1	1	1
82.	Wojskowy Instytut Techniki Pancernej i Samochodowej w Sulejówku	1	1	1
	Razem	383	352	263

Źródło: opracowanie własne.

Ograniczenia w dostępie online

Do blisko 1/4 zasobów internetowych dostęp jest ograniczony. W zdecydowanej większości przypadków, jak wynika z tab. 2, tylko pracownicy danego instytutu lub uprawnione osoby posiadają prawo do korzystania z określonego zasobu. Dotyczy to 88,2% zbiorów danych. Ograniczeniem jest też wymóg rejestracji, spotykany w 8,2% zasobów, ale zazwyczaj zarejestrowany użytkownik otrzymuje dostęp do danych po spełnieniu tego wymogu. Stosunkowo niewielki odsetek, niecałe 5%, baz danych ma płatny charakter.

Ponadto zidentyfikowano dwa rodzaje ograniczeń, które dotyczą zbiorów danych niedostępnych online. Mianowicie, 3 bazy danych są dostępne odpłatnie, a jeden zasób mogą przeszukiwać tylko pracownicy danej jednostki.

Tabela 2. Ograniczenia w dostępie online

Rodzaj ograniczenia	Liczba (n = 85)	Procent**
Dostęp tylko dla pracowników	72	84,7
Tylko uprawnione osoby	3	3,5
Wymóg rejestracji	7	8,2
Płatny dostęp	4	4,7

** Procenty nie sumują się do 100, gdyż w jednej bazie występowały dwa rodzaje ograniczeń – Dostęp tylko dla pracowników i Wymóg rejestracji.

Źródło: opracowanie własne.

Rodzaj danych

Surowe dane badawcze, jak ilustruje tab. 3, są zawarte w mniej niż 40% zbiorów danych opracowywanych przez instytuty. Blisko 58% zbiorów zawiera metadane. W tej grupie mieszczą się m.in. rejestry osób, organizacji, realizowanych projektów, nieruchomości, wyposażenia technicznego, publikacji naukowych pracowników instytutów badawczych. Pełnotekstowe bazy danych stanowią niewiele ponad 9% zasobów dostępnych online[6]. Stosunkowo nieduży odsetek zbiorów stanowią te, które zawierają materiały graficzne.

Prawie 54% (53,6%) zbiorów z metadanymi stanowią źródła o charakterze bibliograficznym rejestrujące opisy różnego typu dokumentów, m.in. publikacji naukowych, norm lub patentów. W przypadku zasobów dostępnych online ten odsetek nie przekracza 52% (51,4%). Same źródła o charakterze bibliograficznym stanowią 30,8% wszystkich zasobów i 30,4% dostępnych online.

Tabela 3. Rodzaj gromadzonych danych

Rodzaj	Wszystkie zasoby – liczba (n = 383)	Procent*	Zasoby dostępne online – liczba (n = 352)	Zasoby dostępne online – procent*
Dane	144	37,6	135	38,4
Metadane	220	57,4	208	59,1
Grafika	25	6,5	18	5,1
Video	1	0,3	0	0,00
Tekst	32	8,37	32	9,1
Brak danych	4	1,0	2	0,6

* Procenty nie sumują się do 100, gdyż niektóre zasoby mogą zawierać kilka rodzajów danych.

Źródło: opracowanie własne.

[6] Jak wspomniano wcześniej, w analizie nie uwzględniono pojedynczych tytułów czasopism naukowych i książek wydawanych i udostępnianych przez instytuty naukowo-badawcze.

Również w grupie zasobów dostępnych online przeważają zbiory metadanych, które stanowią prawie 60% wszystkich zbiorów. Dane badawcze są zawarte w niecałych 39% zbiorów.

Jeśli zestawimy zasoby dostępne online bez ograniczeń, jak ilustruje tab. 4, to odsetek zasobów z surowymi danymi spada do niecałych 25%. Przeważają zbiory z metadanymi, które stanowią blisko 73% analizowanych zasobów. Ponad połowę (54,5%) zbiorów z metadanymi stanowią zasoby o charakterze bibliograficznym.

W przypadku zasobów, do których dostęp w Internecie jest ograniczony, surowe dane, jak wynika z tab. 4, stanowią ponad 80% tego typu zbiorów. Udział metadanych w tej grupie nie przekracza 17%.

Tabela 4. Rodzaj danych – zasoby online dostępne bez ograniczeń i z ograniczeniami

Rodzaj danych	**Zasoby online dostępne bez ograniczeń**		**Zasoby online dostępne z ograniczeniami**	
	Liczba (n = 263)	**Procent***	**Liczba (n = 85)**	**Procent***
Dane	65	24,7	69	81,2
Metadane	191	72,6	14	16,5
Grafika	12	4,6	5	5,9
Tekst	29	11	3	3,5

* Procenty nie sumują się do 100, gdyż niektóre zasoby mogą zawierać kilka rodzajów danych.

Źródło: opracowanie własne.

Format danych

Ogółem w 383 zbiorach informacji, jak ilustruje tab. 5, zidentyfikowano 25 różnych rodzajów formatów danych, w tym 23 w przypadku zbiorów dostępnych online[7]. Każdy z nich, poza formatem flash, można przetwarzać przy pomocy darmowego oprogramowania dostępnego w Internecie, ale problemem pozostaje kwestia wygody pracy z danymi. Najczęściej, w blisko 88% przypadków, dane były udostępniane w formacie html[8]. Stosunkowo wysoki, wynoszący 57,6%, jest jednak odsetek zasobów, które prezentują dane wyłącznie w tym formacie, co zmusza użytkownika do pracochłonnej konwersji do innego formatu, kiedy chce przetwarzać dane np. w arkuszu kalkulacyjnym. W przypadku surowych danych procent zasobów dostępnych tylko html wynosi 52,8, a w przypadku metadanych 70,7.

Znacznie rzadziej dane były dostępne w innych popularnych formatach, takich jak csv, xls, xml lub txt, które znacznie ułatwiają ich dalsze przetwarzanie. Pro-

[7] W przypadku 95 (66 dostępnych online) zbiorów nie udało się ustalić formatu danych. Dotyczy to zwłaszcza zasobów tworzonych przez Instytut Leśnictwa, które są dostępne tylko dla pracowników instytutu. Na stronie WWW instytutu brak informacji o formacie danych.

[8] Wynika to z faktu, że w tym formacie najczęściej są wyświetlane dane wyszukane przez użytkownika.

blemem dla użytkownika jest sytuacja, kiedy dane numeryczne są prezentowane tylko w postaci plików graficznych w formacie jpg, gif, czy png[9]. Odsetek takich zasobów nie jest jednak duży, zaledwie 5,6%.

Tabela 5. Format danych

Format	Ogółem – liczba (n = 288)	Procent	Dostępne online – liczba (n = 286)	Dostępne online – procent
1. HTML	253	87,8	253	88,5
2. PDF	59	20,5	59	20,6
3. XLS	47	16,3	45	15,7
4. TXT	16	5,6	16	5,6
5. RTF	14	4,9	14	4,9
6. XML	10	3,5	10	3,5
7. JPG	9	3,1	9	3,1
8. CSV	7	2,4	7	2,4
9. BIBTEX	6	2,1	6	2,1
10. RDF	5	1,7	5	1,7
11. PNG	6	2,1	6	2,1
12. OAI-PMH	4	1,4	4	1,4
13. RIS	4	1,4	4	1,4
14. DOC	3	1,0	3	1,0
15. GIF	2	0,7	2	0,7
16. SHP	2	0,7	2	0,7
17. DJVU	1	0,3	1	0,3
18. SVG	1	0,3	1	0,3
19. dBase	1	0,3	0	0,0
20. JSON	1	0,3	1	0,3
21. NetCDF	1	0,3	1	0,3
22. TIFF	1	0,3	0	0,0
23. Flash	1	0,3	1	0,3
24. GML	1	0,3	1	0,3
25. GeoMedia MS Access	1	0,3	1	0,3

* Procenty nie sumują się do 100, gdyż dane mogą być prezentowane w kilku formatach.

Źródło: opracowanie własne.

86,7% danych zgromadzono w bazach danych, natomiast 13,3% danych jest dostępnych jedynie w formie statycznych plików, głównie w formacie html lub pdf.

[9] Trzeba jednak zaznaczyć, że w wielu przypadkach dane są prezentowane w kilku formatach, np. w Krajowym Rejestrze Nowotworów (KRN 2015).

Warunki wykorzystania danych

Jak wynika z tab. 6, ponad 87% zbiorów danych dostępnych online, w tym blisko 84% zasobów dostępnych online bez ograniczeń, nie zawiera informacji o warunkach ich wykorzystania. Tylko 19,51% (tab. 7) instytutów udostępniających zasoby dołącza do nich informację na temat zasad ponownego wykorzystania danych.

Tabela 6. Informacja o warunkach wykorzystania danych

Dostępność informacji	**Ogółem**	**Procent**	**Zasoby dostępne online – ogółem (n = 352)**	**Procent**	**Zasoby dostępne online bez ograniczeń – ogółem (n = 263)**	**Procent**
Jest (w tym licencje Creative Commons)	47 (4)	12,27	47 (4)	13,35	43	16,35
Brak informacji	336	87,73	305	86,65	220	83,65

Źródło: opracowanie własne.

Tabela 7. Instytuty udostępniające informację o warunkach wykorzystania danych

Lp.	**Nazwa jednostki**	**Liczba**
1.	Państwowy Instytut Geologiczny. Państwowy Instytut Badawczy	18
2.	Ośrodek Przetwarzania Informacji. Państwowy Instytut Badawczy	7
3.	Instytut Ekologii Terenów Uprzemysłowionych	3
4.	Instytut Włókiennictwa w Łodzi	3
5.	Centralny Instytut Ochrony Pracy. Państwowy Instytut Badawczy	2
6.	Instytut Badań Edukacyjnych	2
7.	Instytut Meteorologii i Gospodarki Wodnej. Państwowy Instytut Badawczy	2
8.	Instytut Ochrony Roślin. Państwowy Instytut Badawczy w Poznaniu	2
9.	Centrum Onkologii – Instytut im. Marii Skłodowskiej-Curie w Warszawie	1
10.	Instytut Geodezji i Kartografii w Warszawie	1
11.	Instytut Logistyki i Magazynowania w Poznaniu	1
12.	Instytut Morski w Gdańsku	1
13.	Instytut Nafty i Gazu. Państwowy Instytut Badawczy w Krakowie	1
14.	Instytut Transportu Samochodowego w Warszawie	1
15.	Instytut Zootechniki. Państwowy Instytut Badawczy w Krakowie	1
16.	Instytut Żywności i Żywienia im. prof. dra med. Aleksandra Szczygła w Warszawie	1
	Razem	47

Źródło: opracowanie własne.

W czterech przypadkach (8,5%) zidentyfikowano zastosowanie licencji typu Creative Commons (CC). Trzy zasoby zostały udostępnione na licencji CC BY, która dopuszcza swobodne pobieranie, przetwarzanie i udostępnianie danych pod warunkiem przywołania autorstwa źródła. W jednym przypadku licencja CC zabraniała komercyjnego wykorzystania danych oraz wymagała rozpowszechniania przetworzonych danych na takiej samej licencji, jak licencja pierwotna.

Najczęściej – 31 razy (66%) – w warunkach występowało zastrzeżenie oznaczenia właściciela zbioru danych w przypadku wykorzystywania go we własnych opracowaniach[10]. Rzadziej, po 19 razy (40,4%), właściciele zasobów pozwalali na tworzenie opracowań opartych na pobranych danych i ich dalsze rozpowszechnianie[11].

W 15 (31,9%) regulaminach zastrzeżono bezpłatne wykorzystanie danych tylko do własnego osobistego użytku. Modyfikacja danych i ich udostępnianie, również komercyjne, wymagało uzyskania zgody właściciela zasobu. Łącznie 18 razy (38,3%) zidentyfikowano zakaz komercyjnego wykorzystywania zasobów (wliczając w to użytek własny) bez zgody właściciela zasobu.

Ponadto, w przypadku dwóch bibliotek cyfrowych swobodne pobieranie, przetwarzanie i udostępnianie danych dotyczyło tylko metadanych obiektów cyfrowych.

Podsumowanie

Analiza wykazała dość wysoki, blisko trzydziestoprocentowy, odsetek instytutów badawczych, które nie udostępniają żadnych danych. Na podstawie informacji z ich stron WWW nie można jednak stwierdzić, że nie tworzą żadnych zasobów. Z dużym prawdopodobieństwem można założyć, że liczba zbiorów danych jest o wiele większa, ale są one dostępnie tylko lokalnie dla potrzeb pracowników danej jednostki. Jednym z powodów takiej sytuacji może być brak odpowiedniej infrastruktury technicznej umożliwiającej gromadzenie i publikowanie danych w Internecie. Taką infrastrukturę stworzyło tylko kilka jednostek, m.in. PIG, OPI czy Główny Instytut Górnictwa[12]. Bez podjęcia dodatkowych badań nie można jednak wyrokować o przyczynach nieudostępniania danych. Poza ograniczeniami finan-

[10] Należy mieć na uwadze, że zasady wykorzystania zasobów mogą obejmować łącznie kilka warunków, np. właściciel zbioru danych pozwala na ich rozpowszechnianie oraz tworzenie opracowań, ale pod warunkiem przywołania informacji o właścicielu danych.

[11] Właściciele badanych zasobów, poza jednym wyjątkiem, nie ograniczali użytkowników w zakresie celu wykorzystania danych. Jedynie Instytut Transportu Drogowego zabraniał wykorzystywania danych „…w żadnym innym celu jak analizy stanu bezpieczeństwa ruchu drogowego" (Regulamin, 2015).

[12] Powyższą tezę może potwierdzać stosunkowo duży odsetek zasobów (ok. 30%) zawierających dane bibliograficzne, które można stosunkowo łatwo udostępnić dzięki istnieniu dość różnorodnej oferty systemów biblioteczno-bibliograficznych, które użytkują biblioteki i ośrodki informacji zlokalizowane w instytutach badawczych. Tymczasem publikacja surowych danych w Internecie wymaga często stworzenia od podstaw nowego systemu komputerowego dostosowanego do specyfiki gromadzonych danych.

sowymi, motywem może być chęć skomercjalizowania wyników badań i obawa przed wykorzystaniem danych przez innych badaczy. Inną przyczyną może być stosunkowo niewielka liczba tego typu danych, która według badaczy nie uzasadnia potrzeby publikowania ich w Internecie.

Wśród zasobów dość duży odsetek zajmują te gromadzące metadane (dane bibliograficzne, opisy osób, organizacji, projektów). Szczególnie dotyczy to zasobów, które są dostępne w Sieci bez ograniczeń. Niewątpliwie mogą być one przedmiotem zainteresowania innych badaczy, ale, jak się wydaje, będą oni bardziej zainteresowani zbiorami zawierającymi surowe dane zgromadzone w trakcie badań. Liczba tego typu zbiorów jest jednak widocznie niższa.

Wprawdzie szeroka oferta darmowego oprogramowania umożliwia użytkownikom przetwarzanie danych zgodnie z własnymi potrzebami, ale w wielu przypadkach wymaga od nich czasochłonnej konwersji danych, zwłaszcza numerycznych, do formatu umożliwiającego przetwarzanie ich w stosowanym oprogramowaniu.

Brak informacji o warunkach wykorzystania danych jest istotną barierą dla innych naukowców chcących wykorzystać je we własnej pracy badawczej. To oraz nakaz wykorzystania danych tylko do własnego osobistego użytku znacząco ogranicza swobodę podejmowania pracy naukowej.

Bibliografia

Babik W. (2008), *Informacja naukowa jako przedmiot zarządzania*, [w:] Zarządzanie informacją w nauce, pod redakcją Diany Pietruch-Reizes, Katowice: Wydawnictwo Uniwersytetu Śląskiego.

Borgman Christine L. (2010), *Research data: who will share what, with whom, when, and why?* (dokument elektroniczny), "RatSWD Working Paper" 161: http://ssrn.com/abstract=1714427 (dostęp: 1.10.2015).

Deklaracja (2005), *Deklaracja Berlińska w sprawie otwartego dostępu do wiedzy w naukach ścisłych i humanistyce* (dokument elektroniczny), (2005), „Biuletyn EBIB" 2 (63): http://www.ebib.pl/2005/63/deklaracja.php (dostęp: 1.10.2015).

DOW (2015), *Definicja Otwartej Wiedzy*, http://opendefinition.org/od/2.0/pl/ (dostęp: 1.10.2015).

EU (2011), *National open access and preservation policies in Europe analysis of a questionnaire to the European Research Area Committee*, Luxembourg: Publications Office of the European Union, http://ec.europa.eu/research/science-society/document_library/pdf_06/open-access-report-2011_en.pdf (dostęp: 1.10.2015).

EU (2012), *Online survey on scientific information in the digital age*, Luxembourg; Taipei, Taiwan: Publications Office of the European Union; European Union Centre in Taiwan, https://ec.europa.eu/research/science-society/document_library/pdf_06/survey-on-scientific-information-digital-age_en.pdf (dostęp: 1.10.2015).

Hofmokl J., Tarkowski A., Bednarek-Michalska B., Siewicz K., Szprot J. (2009), *Przewodnik po otwartej nauce*, Warszawa: Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego Uniwersytetu Warszawskiego.

KRN (2015), *Krajowy Rejestr Nowotworów*, http://onkologia.org.pl/ (dostęp: 1.10.2015).

Leśniak A., Morys-Twarowski M., Siewicz K., Starczewski M., Stępińska-Ustasiak L., Szprot J. (2014), *Otwarta nauka w Polsce 2014. Diagnoza*, red. J. Szprot, Warszawa: Wydawnictwa ICM.

Materska K. (2007), *Informacja w organizacjach społeczeństwa wiedzy*, Warszawa: Stowarzyszenie Bibliotekarzy Polskich.

Molloy J. C. (2011), *The Open Knowledge Foundation: open data means better science*, "PLoS Biol" 9(12): e1001195. doi:10.1371/journal.pbio.1001195.

Murray-Rust P. (2008), *Open Data in Science*, "Serials Review" 34(1): 52–64, doi: 10.1016/j.serrev.2008.01.001.

Murray-Rust P., Neylon C., Pollock R., Wilbanks J. (2010), *Panton Principles, principles for open data in science,* http://pantonprinciples.org/ (dostęp: 1.10.2015).

OECD (2007), *OECD principles and guidelines for access to research data from public funding* (document elektroniczny), http://www.oecd.org/sti/sci-tech/38500813.pdf (dostęp: 1.10.2015).

ORGIB (2014), http://www.rgib.org.pl/index.php?option=com_content&view=article&id=141&Itemid=71 (dostęp: 1.10.2015).

Regulamin (2015), *Regulamin*, https://www.obserwatoriumbrd.pl/pl/p/pobr_pl/rejestracja/regulamin (dostęp: 1.10.2015).

Reichman O. J., Jones Matthew B., Schildhauer Mark P. (2011), *Challenges and opportunities of open data in ecology*, "Science" 331 no. 6018: 703–705, doi: 10.1126/science.1197962.

Sapa R. (2013), *Realizacja funkcji repozytoryjnych przez największe przedsięwzięcia zarejestrowane w Federacji Bibliotek Cyfrowych tworzone i współtworzone przez uczelnie,* „Przegląd Biblioteczny" 2: 117–132.

Tenopir C., Allard S., Douglass K., Aydinoglu A. U., Lei W., Read E., Manoff M., Frame M. (2011), *Data Sharing by Scientists: Practices and Perceptions*, "PLoS ONE" 6(6): e21101, doi:10.1371/journal.pone.0021101.

Uhlir P. F., Schröder P. (2007), *Open data for global science,* "Data Science Journal" 17: 36–53.

Ustawa (2010) *Ustawa z dnia 30 kwietnia 2010 r. o instytutach badawczych*, „Dziennik Ustaw" 2010 nr 96 poz. 618.

Wessels B., Finn R. L., Linde P., Mazzetti P., Nativi S., Riley S., Smallwood R., Taylor M. J., Tsoukala V., Wadhwa K., Wyatt S. (2014), *Issues in the development of open access to research data*, "Prometheus: Critical Studies in Innovation" 32:1, 49–66, doi: 10.1080/08109028.2014.956505.