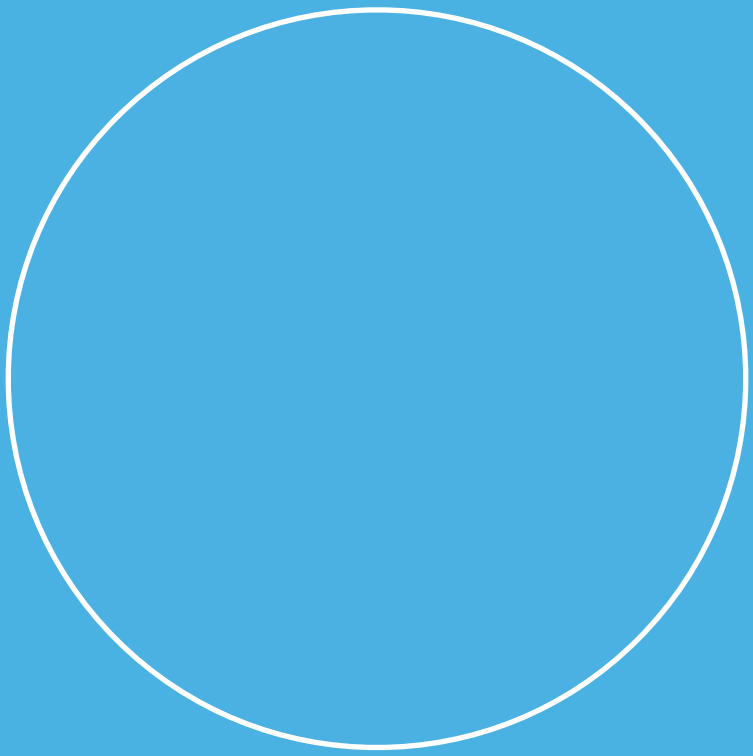
Jak korzystać z zasobów w repozytoriach danych

Wersja 2.0



Opracowanie: Natalia Gruenpeter
Konsultacja: Wojciech Fenrich, dr Krzysztof Siewicz, Jakub Szprot

Pierwsza wersja broszury powstała w ramach projektu Dziedzinowe Repozytoria Otwartych Danych Badawczych, finansowanego ze środków Programu Operacyjnego Polska Cyfrowa i jest dostępna na stronie projektu (https://drodb.icm.edu.pl/materialy-2/).

Uniwersytet Warszawski
Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego
ul. Tyniecka 15/17
02-630 Warszawa
www.icm.edu.pl

UNIWERSYTET WARSZAWSKI

icm

Jak korzystać z zasobów w repozytoriach danych

Wersja 2.0

Spis treści

Słowniczek

Dane badawcze – dane zebrane lub wytworzone jako materiał do analizy w ramach badań naukowych.

Data access committee – składająca się z ekspertów grupa, do której należy decyzja o udostępnieniu zbioru danych.

Data journal – czasopismo naukowe, które publikuje artykuły opisujące zbiory danych badawczych, udostępnione w repozytoriach danych lub (rzadko) w formie suplementu do samego artykułu.

Data management plan – zob. plan zarządzania danymi.

DMP – zob. plan zarządzania danymi.

DOI – ang. „Digital Object Identifier", jeden z trwałych identyfikatorów obiektów cyfrowych, pozwalający na ich odnalezienie w internecie niezależnie od wiodącego do nich adresu URL. Posiadający DOI zbiór danych można za jego pomocą zidentyfikować nawet wtedy, gdy zostanie on przeniesiony na inny serwer czy do innego repozytorium.

Embargo – okres, przez który dane badawcze nie mogą zostać udostępnione publicznie. Jest on zwykle wykorzystywany po to, aby uzyskać związane z nimi patenty i/lub inne prawa własności intelektualnej oraz przygotować oparte na nich publikacje naukowe. Po jego upływie opublikowanie danych badawczych staje się możliwe.

FAIR – akronim słów „findable" (możliwy do znalezienia), „accessible" (dostępny), „interoperable" (interoperacyjny) i „reusable" (możliwy do ponownego wykorzystania), określający wymogi, jakie powinny spełniać udostępnione dane badawcze.

Interoperacyjność – cecha tych danych, które można łączyć z innymi danymi, wykorzystywać w wielu różnych systemach komputerowych i analizować przy użyciu różnorodnego oprogramowania.

Licencja – upoważnienie do korzystania w określony sposób z utworu lub bazy danych. Przedmiotem licencji może być na przykład zbiór danych badawczych.

Licencje Creative Commons – popularne wzory licencji opracowane przez organizację Creative Commons.

Metadane – ustrukturyzowane informacje opisujące zasoby informacji, np. zbiory danych badawczych. Metadane zawierają informacje o formie i treści zasobów, dzięki czemu pozwalają na ich wyszukiwanie i identyfikację oraz zarządzanie nimi.

Ograniczony dostęp – model dostępu, w którym dane udostępniane są jedynie określonym osobom (np. tym, które uzyskały zgodę dysponenta danych) lub kategoriom osób (np. prowadzącym badania naukowe lub zatrudnionym w konkretnej instytucji).

Otwarte dane badawcze – dostępne za pośrednictwem internetu dane badawcze, które można wykorzystywać bez ponoszenia opłat oraz bez istotnych ograniczeń technicznych i prawnych.

Plan zarządzania danymi (*data management plan*, DMP) – dokument opisujący to, co będzie działo się z danymi podczas projektu badawczego i po jego zakończeniu. Ma on charakter „żywego dokumentu", który może i powinien zmieniać się wraz ze zmianami pojawiającymi się w innych obszarach projektu badawczego.

Ponowne wykorzystanie – ogólny termin odnoszący się do technicznych, prawnych i metodologicznych uwarunkowań użycia danych przez dowolne osoby i/lub instytucje, w szczególności te, które nie były zaangażowane w ich wytworzenie.

Repozytorium danych – serwis internetowy służący do deponowania (umieszczania), przechowywania i udostępniania za pośrednictwem internetu danych badawczych w formie cyfrowej.

Wstęp

Niniejsza broszura jest przewodnikiem dla osób chcących korzystać z danych badawczych udostępnianych przez naukowców. Zawarto w niej informacje o tym, gdzie i w jaki sposób szukać takich danych oraz jak korzystać z repozytoriów danych, biorąc pod uwagę warunki dostępu oraz wykorzystania zasobów. Broszura uwzględnia zarówno podstawowe kwestie techniczne, takie jak wyszukiwanie, pobieranie i cytowanie danych, jak i zarys prawnych aspektów korzystania z danych badawczych. Przybliża ponadto konkretne rozwiązania, takie jak czasopisma naukowe publikujące artykuły opisujące zbiory danych, wyszukiwarki zbiorów danych, a także narzędzia prawne umożliwiające szeroki zakres wykorzystania danych (wolne licencje).

Część 1. Sposoby wyszukiwania danych badawczych

Otwarte udostępnianie danych badawczych zyskuje coraz większe znaczenie zarówno jako element polityk otwartości instytucji naukowych, jak i jako rozwiązanie promowane w rozmaitych inicjatywach na rzecz otwartej nauki. Udostępniania danych powiązanych z publikacjami wymagają instytucje finansujące i prowadzące badania naukowe, a coraz częściej także wydawcy naukowi, chcący zapewnić możliwość weryfikacji twierdzeń zawartych w artykułach. Rezultaty badań w postaci danych mogą być replikowane, cytowane i ponownie wykorzystywane przez innych badaczy, a także przez dydaktyków, popularyzatorów wiedzy, dziennikarzy, profesjonalistów z różnych dziedzin i wszystkich zainteresowanych obywateli. Wobec rosnącej ilości danych wytwarzanych i zapisywanych w formie cyfrowej wyzwaniem pozostaje zapewnienie możliwości ich łatwego wyszukiwania. Kluczowe znaczenie mają tutaj sposób opisania danych oraz wybór miejsca ich przechowywania, a najlepszą praktyką, wymaganą przez instytucje finansujące badania, pozostaje deponowanie danych w repozytoriach, czyli specjalnych systemach informatycznych. Powinny one zapewniać bezpieczne i długoterminowe przechowywanie danych, ułatwiać ich wyszukiwanie, pobieranie i cytowanie z wykorzystaniem trwałych identyfikatorów oraz dawać możliwość wyboru odpowiednich licencji.

Repozytoria danych badawczych

Repozytoria danych badawczych, podobnie jak repozytoria publikacji, mogą mieć charakter dziedzinowy lub instytucjonalny. Mogą też łączyć funkcje obu typów, jeżeli prowadząca je instytucja zajmuje się badaniami w określonym obszarze lub prowadzi repozytorium z myślą o konkretnym projekcie badawczym.

Repozytoria ogólnego przeznaczenia

Dane badawcze mogą być deponowane w repozytoriach ogólnego przeznaczenia, pozwalających na udostępnianie danych ze wszystkich dziedzin nauki. Taki charakter mają na przykład repozytoria Zenodo (https://zenodo.org) czy Figshare (https://figshare.com).

RepOD

Repozytorium Otwartych Danych

W Polsce od 2015 roku działa Repozytorium Otwartych Danych (https://repod.icm.edu.pl/) prowadzone przez Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego Uniwersytetu Warszawskiego. Jest to repozytorium ogólnego przeznaczenia służące do otwartego udostępniania każdego typu danych badawczych wytworzonych, zebranych lub opisanych na potrzeby działalności naukowej. Obecna wersja serwisu utworzona została w ramach projektu Dziedzinowe Repozytoria Otwartych Danych Badawczych. Więcej informacji na temat działania serwisu znaleźć można w witrynie informacyjnej (https://repod.icm.edu.pl/info/).

Repozytoria dziedzinowe

Repozytoria dziedzinowe gromadzą dane z konkretnych obszarów wiedzy, co jest istotne z punktu widzenia użytkowników poszukujących danych. Przeszukiwanie właściwego repozytorium zwiększa bowiem prawdopodobieństwo szybkiego znalezienia interesujących danych. Dodatkowo, repozytoria dziedzinowe wdrażają rozwiązania umożliwiające opisanie i udostępnienie danych zgodnie ze standardami stosowanymi w konkretnym obszarze wiedzy, takie jak dziedzinowe schematy metadanych. W rezultacie, dane gromadzone w tego typu repozytoriach zwykle posiadają bogate metadane, które umożliwiają przeszukiwanie zasobów na ich podstawie.

Przykładowe repozytoria dziedzinowe
Archeology Data Services https://archaeologydataservice.ac.uk University of York, archeologia EarthChem http://www.earthchem.org/ Lamont-Doherty Earth Observatory, Columbia University, geochemia, petrologia, geochronologia Repozytorium Danych Społecznych (RDS) https://rds.icm.edu.pl/ Uniwersytet Warszawski, Instytut Filozofii i Socjologii Polskiej Akademii Nauk, nauki społeczne Macromolecular Xtallography Raw Data Repository (MX-RDR) https://mxrdr.icm.edu.pl/ Uniwersytet Warszawski, Uniwersytet im. Adama Mickiewicza w Poznaniu, krystalografia Open Forest Data https://openforestdata.pl/ Instytut Biologii Ssaków Polskiej Akademii Nauk, Politechnika Białostocka, nauki przyrodnicze

Repozytorium Danych Społecznych (RDS) i Macromolecular Xtallography Raw Data Repository (MX-RDR) to dwa repozytoria dziedzinowe utworzone w ramach projektu Dziedzinowe Repozytoria Otwartych Danych Badawczych realizowanego przez Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego Uniwersytetu Warszawskiego we współpracy z Instytutem Studiów Społecznych im. Prof. Roberta Zajonca Uniwersytetu Warszawskiego, Instytutem Filozofii i Socjologii Polskiej Akademii Nauk i Uniwersytetem im. Adama Mickiewicza w Poznaniu.

RDS
Repozytorium
Danych Społecznych

Repozytorium Danych Społecznych (https://rds.icm.edu.pl/) służy do udostępniania wszelkiego typu danych społecznych, jakościowych oraz ilościowych. Serwis umożliwia korzystanie z rozwiązań dostosowanych do tego typu danych, takich jak słowniki czy schemat metadanych dla nauk społecznych. Zaimplementowane w repozytorium metadane obejmują m.in. informacje metodologiczne o podstawowej jednostce analizy lub obserwacji, badanej populacji, procedurze doboru próby czy technice wykorzystanej do zbierania danych. Metadane te stanowią także osobne pola, z których skorzystać można przy zaawansowanym wyszukiwaniu danych. W Repozytorium Danych Społecznych działają obecnie kolekcje Archiwum Danych Jakościowych (ADJ) prowadzonego przez Instytut Filozofii i Socjologii PAN oraz Polskiego Archiwum Danych Społecznych (PADS), wspólnego przedsięwzięcia Uniwersytetu Warszawskiego i Instytutu Filozofii i Socjologii PAN. Dodatkowe informacje znaleźć można w witrynach informacyjnych (https://adj.ifispan.pl/ oraz https://pads.org.pl/).

MX-RDR
Macromolecular Xtallography
Raw Data Repository

Macromolecular Xtallography Raw Data Repository (https://mxrdr.icm.edu.pl/) służy do archiwizacji i udostępniania surowych danych dyfrakcyjnych zarejestrowanych dla kryształów makromolekuł. Dane te rejestrowane są głównie na liniach krystalograficznych w dużych ośrodkach synchrotronowych lub przy wykorzystaniu dyfraktometrów i innych źródeł promieniowania rentgenowskiego. W repozytorium można udostępniać

wszystkie zestawy dyfrakcyjne, nawet takie, które z różnych względów nie były wykorzystane do badań, a powinny być zachowane. Serwis umożliwia automatyczne uzupełnianie kluczowych metadanych krystalograficznych na podstawie nagłówków obrazów dyfrakcyjnych, informacji zawartych w bazie PDB lub plików cif oraz XDS. Więcej informacji znaleźć można w witrynie informacyjnej (https://info.mxrdr.icm.edu.pl/).

Repozytoria instytucjonalne

Repozytoria instytucjonalne gromadzą dane badawcze wytworzone i opracowane przez osoby afiliowane przy konkretnej instytucji lub przez grantobiorców realizujących finansowane przez nią badania. Dla naukowców szukających danych mogą być przydatne zwłaszcza wtedy, gdy prowadząca repozytorium instytucja specjalizuje się w określonym obszarze.

Przykładowe repozytoria instytucjonalne
DataCat: The Research Data Catalogue http://datacat.liverpool.ac.uk University of Liverpool Open Data LMU https://data.ub.uni-muenchen.de Ludwig-Maximilians-Universität München CaltechDATA https://data.caltech.edu/ California Technical Institute of Technology

Funkcje repozytoriów instytucjonalnych mogą pełnić także kolekcje w repozytoriach opartych na oprogramowaniu Dataverse, jak w norweskim serwisie DataverseNO (https://dataverse.no/) prowadzonym przez Arctic University of Norway czy niderlandzkim DataverseNL (https://dataverse.nl/) prowadzonym przez Data Archiving and Network Services (DANS).

W Polsce rozwiązanie takie oferuje Repozytorium Otwartych Danych. Założenie kolekcji instytucjonalnej w RepOD umożliwia gromadzenie w jednym miejscu danych badawczych wytwarzanych w ramach badań naukowych prowadzonych w danej instytucji oraz udostępnianie ich zgodnie ze światowymi standardami, między innymi zasadami FAIR. Instytucja może sprawować nadzór merytoryczny nad swoją kolekcją, współpracując przy jej prowadzeniu z ICM UW. Więcej informacji znaleźć można w witrynie informacyjnej repozytorium (w zakładce Kolekcje instytucjonalne, https://repod.icm.edu.pl/info/?page_id=429).

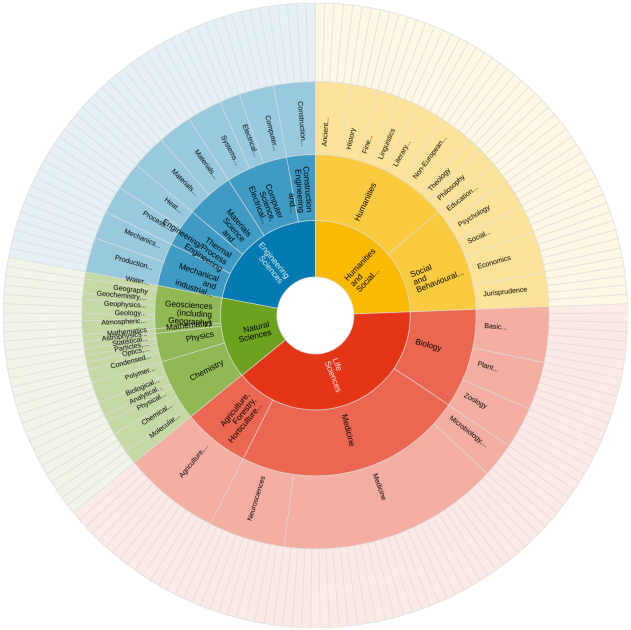
Rejestry repozytoriów

Poszukując właściwego repozytorium, zarówno do udostępnienia, jak i wyszukiwania danych badawczych, warto skorzystać z rejestrów repozytoriów.

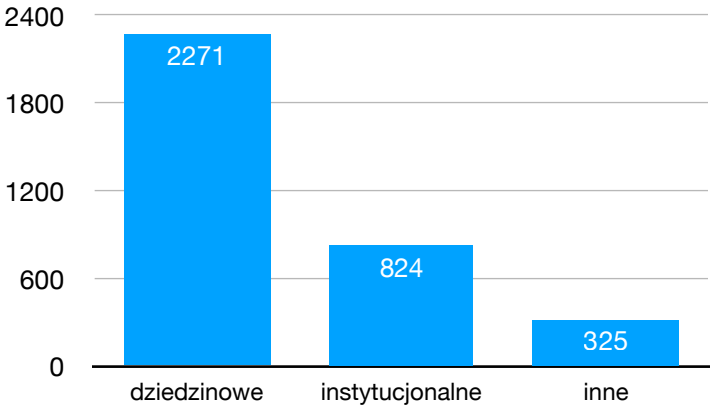
Registry of Research Data Repositories (re3data.org)

Najpełniejszym źródłem wiedzy na temat repozytoriów danych badawczych jest obecnie międzynarodowy serwis Registry of Research Data Repositories (re3data.org), w którym znajduje się ich prawie 3000 (dane z listopada 2022 roku). Zasoby tego rejestru można przeszukiwać według wpisanych słów lub przeglądać, korzystając z trzech podstawowych kryteriów: kraju, obszaru wiedzy oraz typu danych. Wyniki wyszukiwania lub przeglądania można następnie zawęzić za pomocą blisko 30 filtrów, uwzględniających m.in. kwestie prawne (np. warunki dostępu i korzystania z repozytorium oraz z danych), instytucjonalne (m.in. rodzaj instytucji prowadzącej repozytorium), techniczne (jak np. oprogramowanie repozytoryjne, rodzaj API) czy zakres przedmiotowy (przedmiot, słowa kluczowe).

Większość zarejestrowanych w bazie re3data.org repozytoriów ma charakter dziedzinowy. Aby znaleźć repozytorium właściwe dla konkretnego obszaru wiedzy, można skorzystać z odpowiednich kryteriów wyszukiwania bądź z interaktywnego koła, uwzględniającego różne poziomy specjalizacji badań.



Obszary wiedzy w rejestrze repozytoriów danych re3data.org
Źródło: Registry of Research Data Repositories (CC BY 4.0)
https://www.re3data.org/browse/by-subject/



Rodzaje repozytoriów w re3data.org
– opracowanie własne na podstawie danych z rejestru (listopad 2022 r.)

Drugą pod względem wielkości kategorią są repozytoria instytucjonalne, które również – ze względu na profil prowadzących je instytucji – mogą mieć charakter zbliżony do repozytoriów dziedzinowych. W kategorii „innych repozytoriów" znajduje się zaś ponad 300 serwisów, wśród których są np. repozytoria prowadzone przez sieci i projekty badawcze, a także serwisy z otwartymi danymi publicznymi, udostępnianymi przez państwowe urzędy i instytucje. Zostały one ujęte w rejestrze, ponieważ informacje sektora publicznego również mogą być wykorzystywane w badaniach naukowych.

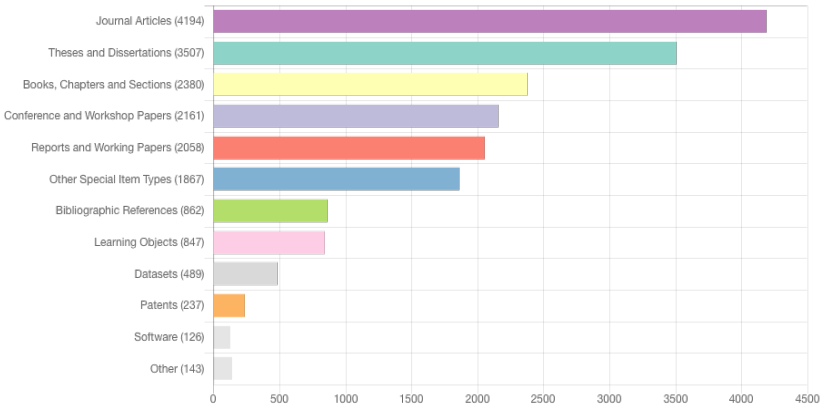
W Polsce dane publiczne udostępniane są za pomocą Centralnego Repozytorium Informacji Publicznej (https://dane.gov.pl), które umożliwia bezpłatny dostęp do informacji z różnych kategorii, takich jak oświata, budżet i finanse, środowisko czy społeczeństwo. Portal gromadzi dane publiczne, które mają potencjał dla rozwoju społeczeństwa informacyjnego i dalszego wykorzystywania, także w nauce. Zbiory danych można przeszukiwać według kategorii, dostawców danych, słów kluczowych, a także wpisując frazę w wyszukiwarce. Wyniki wyszukiwania można dodatkowo filtrować, aby dotrzeć do interesujących danych.

Directory of Open Access Repositories (OpenDOAR)

Rejestr OpenDOAR (https://v2.sherpa.ac.uk/opendoar/) obejmuje wszelkiego typu otwarte repozytoria służące do przechowywania materiałów o charakterze naukowym. W serwisie można znaleźć repozytoria gromadzące m.in. artykuły naukowe, monografie, publikacje pokonferencyjne, raporty, dane badawcze, oprogramowanie, a także materiały archiwalne, audiowizualne czy dydaktyczne. Aktualnie w OpenDOAR zarejestrowanych jest ponad 5970 serwisów, z czego prawie 490 przechowuje dane badawcze (stan na listopad 2022 roku).

Co do zasady repozytoria wpisane do OpenDOAR oferują otwarty dostęp do zasobów, jednakże niektóre zdeponowane w nich treści mogą być objęte ograniczeniami dostępu, np. embargiem. Aby znaleźć repozytoria danych badawczych można skorzystać z zaawansowanego

wyszukiwania, które pozwala m.in. określić rodzaj repozytorium, zastosowane w nim oprogramowanie, typ zasobów czy zakres dziedzinowy.



Typy zasobów w repozytoriach zarejestrowanych w OpenDOAR (listopad 2022 r.)
Źródło: OpenDOAR Statistics, Content Types Overview (CC BY-NC-ND 4.0)
https://v2.sherpa.ac.uk/view/repository_visualisations/1.html

Wyszukiwarki danych badawczych

Stosunkowo nowymi narzędziami są wyszukiwarki danych badawczych, które – podobnie jak inne wyszukiwarki internetowe – nie przechowują danych, tylko indeksują zasoby innych serwisów.

DataCite Search

DataCite Search (https://search.datacite.org/) to serwis służący do wyszukiwania danych badawczych, prowadzony przez organizację DataCite, która dostarcza trwałe identyfikatory DOI (Digital Object Identifier) dla zbiorów danych oraz innych rezultatów badań. Serwis gromadzi metadane obiektów poprzez identyfikatory DOI, które odsyłają do zasobów z ponad 2600 źródeł – repozytoriów prowadzonych przez różnego typu instytucje. Z jego pomocą można znaleźć różne wyniki badań, ale użytkownicy poszukujący danych badawczych mogą zaznaczyć opcję „zbiór danych" (*dataset*), zawężając w ten sposób

wyniki wyszukiwania. Aktualnie baza DataCite zawiera metadane ponad 34 mln prac, z czego ponad 13 mln to zbiory danych (dane z listopada 2022 r.). Nie wszystkie zasoby są otwarte, a wyszukiwarka nie umożliwia niestety filtrowania wyników pod tym kątem. Serwis ułatwia natomiast cytowanie danych, ponieważ generuje rekordy bibliograficzne w ośmiu formatach: APA, Harvard, MLA, Vancouver, Chicago, IEEE, BibTex, RIS.

OpenAIRE Explore

OpenAIRE (https://www.openaire.eu/) to europejska organizacja rozwijająca infrastrukturę otwartej nauki, która agreguje i łączy informacje na temat rezultatów badań z informacjami o projektach badawczych, źródłach danych, serwisach czy instytucjach naukowych. Moduł OpenAIRE Explore (https://explore.openaire.eu/) umożliwia wyszukiwanie różnego typu informacji, w tym danych badawczych. Aby znaleźć dane należy wybrać w polu wyszukiwania „Research outcomes", a następnie zaznaczyć opcję „Research data" bądź skorzystać z wyszukiwania zaawansowanego. Wyniki można przefiltrować, korzystając z pól określających m.in. warunki dostępu, zakres czasowy czy typ danych. System OpenAIRE harwestuje metadane z różnego typu źródeł, takich jak czasopisma naukowe, repozytoria, archiwa, systemy CRIS czy agregatory.

Google Dataset Search

Google Dataset Search (https://datasetsearch.research.google.com/) to serwis Google służący do wyszukiwania zbiorów danych. Informacje zbierane są na podstawie metadanych udostępnianych przez repozytoria, wzbogacanych następnie o inne indeksowane przez Google dane, zaczerpnięte m.in. z Google Scholar. Serwis umożliwia wyszukiwanie informacji o zbiorach powszechnie dostępnych danych gromadzonych w repozytoriach. Wyniki przefiltrować można pod kątem ostatniej aktualizacji, rodzaju danych, prawa do użytkowania i zakresu tematycznego. Wyszukiwarka ułatwia też cytowanie zbioru i dalsze udostępnianie informacji na jego temat (np. w mediach społecznościowych).

Data journals

Jednym z rozwiązań mających na celu zachęcanie badaczy do otwartego udostępniania danych są *data journals*, czyli recenzowane czasopisma z artykułami na temat zbiorów danych badawczych. Publikowane w nich artykuły (*data descriptors*) mogą zawierać załączniki z omawianymi danymi lub wskazywać miejsca przechowywania tych danych – zwykle repozytoria. W zależności od dziedziny nauki lub przyjętych przez redakcję wytycznych publikacje mogą ponadto zawierać takie elementy, jak: ilościowe i jakościowe informacje na temat zbiorów danych, opis ich źródła lub metodologii ich pozyskiwania, opis ich znaczenia, a także informacje na temat warunków korzystania z danych (rodzaju licencji). Zwykle *data journals* udostępniają też listy rekomendowanych repozytoriów, z których korzystać mogą naukowcy publikujący dane badawcze. Listy te mogą być pomocne również w wyszukiwaniu danych, zwłaszcza kiedy czasopismo specjalizuje się w określonym obszarze badawczym.

Przykładowe *data journals*
Data in Brief https://www.sciencedirect.com/journal/data-in-brief Elsevier Scientific Data https://www.nature.com/sdata/ Springer Nature Journal of Open Archaeology Data https://openarchaeologydata.metajnl.com/ Ubiquity Press Research Data Journal for the Humanities and Social Sciences https://brill.com/view/journals/rdj/rdj-overview.xml Brill Geoscience Data Journal https://rmets.onlinelibrary.wiley.com/journal/20496060 Wiley, Royal Meteorological Society

Część 2. Najpopularniejsze funkcjonalności repozytoriów

Wyszukiwanie i pobieranie

Repozytoria danych badawczych umożliwiają użytkownikom zarówno przeszukiwanie (*search*), jak i przeglądanie (*browse*) zasobów. Wyniki obydwu działań można dodatkowo zawęzić, korzystając z oferowanych przez repozytoria filtrów. Najczęstsze to opcja przeszukiwania lub przeglądania według:

a) możliwości uzyskania dostępu do danych (*terms of access* lub *access rights*);
b) warunków korzystania z danych, ujętych zwykle w formule licencji (*terms of use*);
c) obszarów wiedzy oraz dziedzin i dyscyplin naukowych (w repozytoriach ogólnego przeznaczenia) bądź słów kluczowych;
d) rodzaju danych;
e) formatu danych;
f) źródła danych (informacje o autorach, zespołach czy instytucjach finansujących i prowadzących badania).

Dostęp do danych badawczych realizowany jest poprzez zapewnienie użytkownikom możliwości pobrania plików na dysk komputera. Dodatkową opcją usprawniającą proces wyszukiwania danych jest funkcja widoku czy podglądu (*preview*, *explore*), umożliwiająca zapoznanie się z zawartością pliku w przeglądarce, bez konieczności jego pobierania. Dobrą praktyką stosowaną w większości repozytoriów jest interfejs API – umożliwiający dostęp maszynowy – oraz możliwość pobierania metadanych przez OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting), co pozwala na automatyczne ich pobieranie i gromadzenie w ramach innych serwisów.

Udostępnianie i cytowanie

Użytkownikom, którzy znaleźli interesujące ich dane i chcą z nich skorzystać, repozytoria oferują narzędzia ułatwiające ich cytowanie, dzielenie się nimi w mediach społecznościowych czy zamieszczanie ich na stronach internetowych. Sugerowana formuła cytowania zwykle dostępna jest na stronie z opisem danych. Wystarczy ją skopiować i wkleić do dowolnego dokumentu, a w razie potrzeby – dostosować do przyjętego sposobu formatowania przypisów bądź bibliografii. Badacze korzystający z narzędzi do zarządzania bazami bibliograficznymi często mogą także skorzystać z opcji wygenerowania rekordu bibliograficznego w jednym z popularnych formatów, takich jak EndNote XML, RIS czy BibTex. Repozytoria dają ponadto możliwość szybkiego i bezpośredniego udostępnienia plików w mediach społecznościowych bądź bezpośredniego osadzenia ich na stronie internetowej poprzez wykorzystanie fragmentu kodu źródłowego (opcja *embed* – osadź).

Kontrola wersji

Funkcjonalnością ważną ze względów merytorycznych jest kontrola zamieszczanych w repozytorium wersji zbiorów danych badawczych, która polega na odnotowaniu wszelkich zmian wraz z datą ich wprowadzenia. To istotna funkcja, ponieważ dane lub opisujące je metadane mogą w toku badań ulegać zmianom (np. narastać). Naukowcy mogą też udostępniać kolejne, coraz lepiej opracowane wersje zbioru, które mają zastąpić poprzednie wersje. Nie znaczy to jednak, że poprzednie wersje powinny tym samym zniknąć, gdyć mogły one już zostać zacytowane lub wykorzystane przez innych badaczy do realizacji ich własnych celów badawczych. Korzystając z danych zamieszczonych w repozytorium, warto zapoznać się z polem „wersje" (*versions*), a cytując dane, należy wskazać odpowiednią.

Funkcje repozytoriów uruchomionych w ramach projektu Dziedzinowe Repozytoria Otwartych Danych Badawczych

RepOD
Repozytorium Otwartych Danych

RDS
Repozytorium
Danych Społecznych

MX-RDR
Macromolecular Xtallography
Raw Data Repository

Repozytoria utworzone w ramach projektu oparte są na Dataverse (https://dataverse.org/), czyli oprogramowaniu *open source* tworzonym i rozwijanym na Uniwersytecie Harvarda. Oferuje ono szereg funkcji pozwalających na przechowywanie, opisywanie oraz udostępnianie danych badawczych w sposób, który ułatwia spełnienie zasad FAIR oraz uwzględnia specyfikę dziedzinową, na przykład odpowiednią konfigurację schematów metadanych.

Wyszukiwanie i pobieranie

Podstawowe filtry wyszukiwania widoczne na stronach głównych repozytoriów uwzględniają możliwość wyboru zakresu wyszukiwania (na poziomie kolekcji, zbiorów danych bądź plików), a także roku, autora, obszaru tematycznego i słów kluczowych. Wyszukiwanie zaawansowane uwzględnia wszystkie pola metadanych serwisu. Pliki dostępne publicznie pobierać można poprzez interfejs graficzny („z przeglądarki") oraz przez API. Jeśli zbiór zawiera wiele plików, można je wszystkie łatwo pobrać jako jedno archiwum ZIP. W wy-

padku plików tabelarycznych istnieje możliwość ich pobrania w kilku tworzonych automatycznie formatach.

Cytowanie

Odnoszenie się do zbiorów danych zdeponowanych w repozytoriach jest łatwiejsze dzięki sugerowanej formule cytowania, która znajduje się tuż pod nazwą zbioru. Funkcja „Cytuj zbiór danych" umożliwia ponadto wygenerowanie rekordu bibliograficznego w formatach EndNote XML, RIS i BibTex, a „Udostępnij" — udostępnienie informacji o danych w popularnych serwisach społecznościowych.

Kontrola wersji

Wersje widoczne są w sugerowanej formule cytowania oraz w zakładce „Wersje" poniżej podstawowych metadanych. Zawiera ona informacje na temat zmian oraz linki do wszystkich wersji. Zmiana w metadanych może zostać opublikowana jako „mała" (1.1; 1.2 itd.) lub „duża" (2.0; 3.0 itd.) wersja zbioru. Zmiana w danych (plikach) zawsze publikowana jest jako „duża" wersja. Poprzednie wersje zbioru pozostają dostępne – być może ktoś już użył tych danych lub je zacytował.

Trwałe identyfikatory

Wszystkie zbiory danych deponowane w repozytoriach otrzymują numery DOI (Digital Object Identifier), ułatwiające wyszukiwanie, cytowanie i jednoznaczną identyfikację określonego zasobu. Osoby deponujące dane badawcze mogą wskazać swój identyfikator ORCID (Open Researcher and Contributor ID), co ułatwia odbiorcom zapoznanie się z dorobkiem naukowym autorów, a także ROR (Research Organization Registry) identyfikujący instytucję naukową.

Powiązane publikacje i zbiory danych

W repozytoriach możliwe jest dodanie informacji na temat publikacji bądź innych zbiorów danych powiązanych z deponowanym zasobem, a także wskazanie przypisanych im trwałych indentyfikatorów. Ułatwia to wyszukiwanie dodatkowych materiałów powiązanych z danymi badawczymi.

Część 3. Dostęp do danych

Dostęp do repozytorium a dostęp do danych badawczych

Rejestr repozytoriów re3data.org rozróżnia dostęp do repozytorium, czyli do bazy (*database access*) oraz dostęp do przechowywanych w repozytorium danych, czyli do zawartości bazy (*data access*). Pierwsze kryterium pozwala wyróżnić trzy rodzaje repozytoriów:

a) otwarte – taki charakter ma zdecydowana większość z nich, dostęp do nich nie jest w żaden sposób ograniczony, a umieszczone tam zasoby przeglądać może każdy,
b) zamknięte – dostęp do nich ma ograniczona grupa osób, zwykle pracownicy danej instytucji, osoby należące do danego zespołu, stowarzyszenia czy sieci badawczej,
c) z ograniczonym dostępem – dostęp może wiązać się z koniecznością rejestracji w portalu, wniesienia opłaty bądź z ograniczeniami innego typu, na przykład z koniecznością wypełnienia wniosku o udzielenie dostępu, przejścia procesu autoryzacji czy szkolenia (może to dotyczyć w szczególności repozytoriów przechowujących dane zebrane w toku badań klinicznych).

Repozytoria oznaczone w rejestrze jako otwarte umożliwiają użytkownikom przeglądanie zasobów, jednak nie oznacza to, że wszystkie zdeponowane w nich zbiory danych są dostępne w sposób otwarty. W podobny sposób zdefiniowane są kryteria w rejestrze OpenDOAR (https://v2.sherpa.ac.uk/opendoar/), który co do zasady zawiera otwarte repozytoria, ale nie wymaga, aby wszystkie zasoby repozytoriów były dostępne w ten sposób.

Ograniczenia w dostępie do danych

Większość repozytoriów umożliwia ustanowienie embarga, czyli okresu, w którym dane po ich zdeponowaniu w repozytorium pozostają

zamknięte. Niezależnie od embarga, przeglądający zasoby użytkownicy mogą zapoznać się z opisem danych (metadanymi) oraz zidentyfikować badaczy, którzy zamieścili je w repozytorium (lub instytucję, która to zrobiła). W takiej sytuacji można skontaktować się z osobami odpowiedzialnymi za dane, aby dowiedzieć się, czy istnieje możliwość zapoznania się z nimi przed upływem okresu embarga.

Inne możliwe ograniczenia w dostępie do danych to m.in:

- konieczność wniesienia opłaty za dostęp,
- wymóg rejestracji w serwisie, czasami połączony z autoryzacją,
- uwierzytelnienie, np. konieczność potwierdzenia tożsamości bądź afiliacji użytkownika chcącego zapoznać się z danymi.

W takich sytuacjach na stronie z opisem zbioru danych znajduje się zwykle informacja o możliwości uzyskania dostępu, np. odnośnik do formularza kontaktowego lub kontakt do osób odpowiedzialnych za udostępnianie danych (*data access committee*).

Część 4. Korzystanie z danych

Prawne aspekty korzystania z danych badawczych

Uzyskanie dostępu do danych nie jest jednoznaczne z możliwością wykorzystania ich w dowolny sposób. Korzystanie z nich musi odbywać się zgodnie z przypisanymi im ograniczeniami prawnymi. Na przykład z danych udostępnionych bez wskazania konkretnej licencji bądź udostępnionych z wyraźnym zastrzeżeniem praw można korzystać jedynie w granicach swobód określonych prawem.

Chodzi tu przede wszystkim o przepisy prawa autorskiego oraz ustawy o ochronie baz danych – dane badawcze mogą podlegać prawu autorskiemu, prawu *sui generis* do baz danych bądź obu tym reżimom jednocześnie. Konsekwencją takiej ochrony jest wąski zakres dopuszczalnego ponownego wykorzystania (dozwolony użytek i jego okrojony odpowiednik w ustawie o ochronie baz danych, obejmujący użytek osobisty, użytek w celach dydaktycznych lub badawczych oraz cele państwowe: bezpieczeństwo wewnętrzne oraz postępowania sądowe lub administracyjne).

Wolne licencje

Z danych udostępnionych na określonej licencji można korzystać zgodnie z zakresem swobód określonych w tej licencji. Istotne są w tym kontekście zwłaszcza wolne licencje. Są one nieodpłatne, pozwalają na bardzo szeroki zakres wykorzystania i nie nakładają na licencjobiorców zobowiązań idących dalej niż obowiązki przekazywania odbiorcom określonych informacji (np. o autorze, źródle, licencji – tzw. klauzule uznania autorstwa) lub obowiązki stosowania takiej samej licencji w przypadku rozpowszechniania modyfikacji utworu (tzw. klauzule *copyleft* lub *sharealike*).

Wśród najbardziej popularnych licencji Creative Commons, wolne licencje to:

1. Creative Commons Uznanie autorstwa (CC-BY) – licencja ta pozwala na praktycznie dowolne wykorzystywanie objętego nią materiału, wymaga od użytkowników jedynie zachowania oznaczeń autorstwa, poszanowania innych praw osobistych (np. zakaz przypisywania sobie poparcia autora dla określonego wykorzystania utworu) oraz przekazania dalej informacji o licencji.
2. Creative Commons Uznanie autorstwa – Na tych samych warunkach (CC-BY-SA) – w porównaniu do licencji CC-BY licencja ta zawiera jedno dodatkowe wymaganie, a mianowicie zobowiązuje użytkownika do opublikowania na takiej samej licencji jego własnych modyfikacji oryginalnego materiału, jeżeli zdecyduje się on je rozpowszechniać.
3. CC0 – wzorzec oświadczenia, na podstawie którego uprawniony zrzeka się w zasadzie wszelkich uprawnień, jakie mogłyby mu przysługiwać w odniesieniu do udostępnianego materiału. Na wypadek, gdyby takie zrzeczenie było nieskuteczne, we wzorcu przewidziano udzielenie szerokiej, nieodpłatnej licencji, bez żadnych wymagań dla użytkownika. Udostępniający zobowiązuje się nie egzekwować przysługujących mu uprawnień.

Warto podkreślić, że istnieją także licencje Creative Commons zawierające ograniczenia korzystania komercyjnego (klauzula NC — *Non-Commercial*) lub zakaz modyfikacji (klauzula ND — *No Derivatives*). Licencje te nie są wolnymi licencjami. Choć nie doprowadzą one do otwarcia danych zgodnie z przyjętym rozumieniem tej otwartości (wymagającym braku takich ograniczeń), dają jednak użytkownikowi większą swobodę niż ta, która wynika z samych przepisów ustawy.

Poza ustaleniem tego, na jakich zasadach udostępniony jest określony zbiór danych, z praktycznego punktu widzenia istotne jest jeszcze sprawdzenie dwóch kwestii. Po pierwsze, czy dane zostały udostępnione na określonej licencji przez osobę uprawnioną. Po drugie, jakiego elementu zbioru dokładnie dotyczy określony zestaw wymagań (ustalenie zakresu licencji).

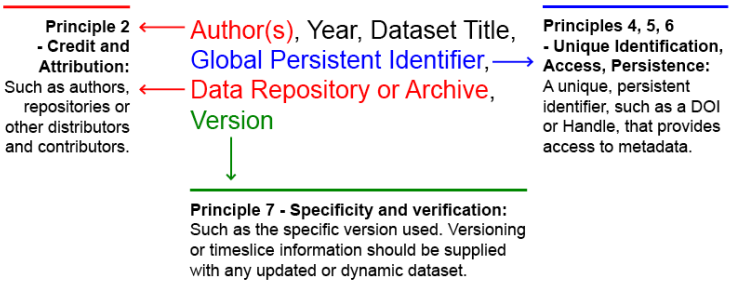
Prawa do danych powstają samoistnie i nie ma publicznych rejestrów, w których można sprawdzić, komu one przysługują. Z kolei umowy przenoszące te prawa są zazwyczaj dostępne tylko stronom. W związku z tym przeciętny użytkownik ma bardzo niewielkie możliwości sprawdzenia, czy udostępniający jest do tego uprawniony. Opiera się to przede wszystkim na zaufaniu. Praktyczna realizacja kwestii pierwszej sprowadza się zatem do analizy wiarygodności samego repozytorium, ale też konkretnego udostępnionego w nim zestawu danych. Zbyt ogólnikowe metadane (w tym zbyt ogólne wskazanie warunków na jakich dane zostały udostępnione) to jedna z rzeczy, które powinny wzbudzić czujność prawną użytkownika i skłonić go do dokładniejszego upewnienia się, czy dane zostały udostępnione legalnie, a udostępniająca je osoba świadomie wybrała określoną licencję.

Natomiast ustalenie zakresu uzyskanej licencji wymaga przede wszystkim uświadomienia sobie, że prawna ochrona danych może dotyczyć zarówno ich zestawu (zbioru) jako całości, jak i jego poszczególnych elementów. Na tych różnych poziomach prawa mogą przysługiwać innym osobom (prawa *sui generis* do bazy jako całości przysługiwać będą producentowi, czyli np. jednostce naukowej, podczas gdy np. prawa do poszczególnych zawartych w bazie fotografii, wywiadów itd. mogą przysługiwać poszczególnym naukowcom lub nawet osobom trzecim). Objęcie niektórych z tych praw (np. dóbr osobistych lub danych osobowych osób badanych, o których informacje znajdują się w zestawie danych) wolną licencją może w ogóle nie być możliwe. Z kolei w wypadku innych elementów lub praw będzie to po prostu wymagać dokładniejszego opisania w metadanych przez samego udostępniającego (powinien on doprecyzować, czy udzielana licencja dotyczy tylko zbioru, czy także poszczególnych jego elementów).

Cytowanie danych

Dane badawcze, które zostały wykorzystane bądź przywołane w publikacjach, należy odpowiednio zacytować, stosując się do dobrych

praktyk i zasad rzetelności naukowej. Ogólne zasady cytowania danych określone zostały w dokumencie „Joint Declaration of Data Citation Principles" (https://www.force11.org/datacitationprinciples), zgodnie z którym dane badawcze uznaje się za pełnoprawne rezultaty badań. Cytowanie danych pozwala na szybką i jednoznaczną identyfikację ich źródła, ułatwia weryfikację twierdzeń zawartych w publikacjach, a także sprzyja ponownemu wykorzystaniu danych. Powinno ponadto ułatwiać dostęp do danych wraz z powiązanymi z nimi metadanymi, dokumentacją, kodem i innymi materiałami niezbędnymi do rzetelnego korzystania z danych. Informacje, które należy uwzględnić w cytowaniu to: autor lub autorzy, tytuł zbioru, rok i miejsce udostępnienia, np. nazwa repozytorium lub archiwum, wersja, trwały identyfikator.



Dataverse Project — Example of a data citation based on the Joint Declaration of Data Citation Principles
Źródło: https://dataverse.org/best-practices/data-citation

Podsumowanie

Korzystając z udostępnionych w repozytoriach danych badawczych należy uwzględnić zarówno aspekty techniczne, np. formaty czy wielkość plików, jak i prawne, związane z warunkami dostępu oraz korzystania z danych. W obydwu tych kwestiach możliwość ponownego wykorzystania danych badawczych zależy od tego, w jaki sposób zbiory przygotowane zostały przez udostępniających. Stosowanie standardowych rozwiązań, w szczególności zgodnych z zasadami FAIR, nie tylko ułatwia wyszukanie danych, lecz także świadczy o ich rzetelnym przygotowaniu, umożliwia jednoznaczne ustalenie źródła danych oraz odpowiednie ich zacytowanie. To z kolei jest ważne z punktu widzenia badacza chcącego skorzystać z danych badawczych w sposób uczciwy i rzetelny.