DANE BADAWCZE

Biblioteka Uniwersytecka w Warszawie

opracowały
Anna Książczak-Gronowska,
Maja Bogajczyk

Warszawa 2020

DANE BADAWCZE

OPRACOWANIE:

MAJA BOGAJCZYK
ANNA KSIĄŻCZAK-GRONOWSKA

Oddział Usług Informacyjnych i Szkoleń Biblioteka Uniwersytecka w Warszawie
ul. Dobra 56/66, 00-312 Warszawa
tel. (22) 55 25 159; (22) 55 25 178; (22) 55 25 179
informator.dziedzinowy@uw.edu.pl
buw.uw.edu.pl

Wydanie 1

Ilustracje: Freepik.com - https://stories.freepik.com/

SPIS TREŚCI

DANE BADAWCZE

Dane badawcze – dane zebrane lub wytworzone jako materiał do analizy w ramach badań naukowych. Dostęp do nich pozwala zweryfikować wyniki zaprezentowane w publikacji naukowej.

Komisja Europejska definiuje dane badawcze jako *informacje, w szczególności fakty, liczby, zebrane do analizy i uważane za podstawę do dalszego wnioskowania, dyskusji lub obliczeń.* ("Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020")

RODZAJE DANYCH BADAWCZYCH

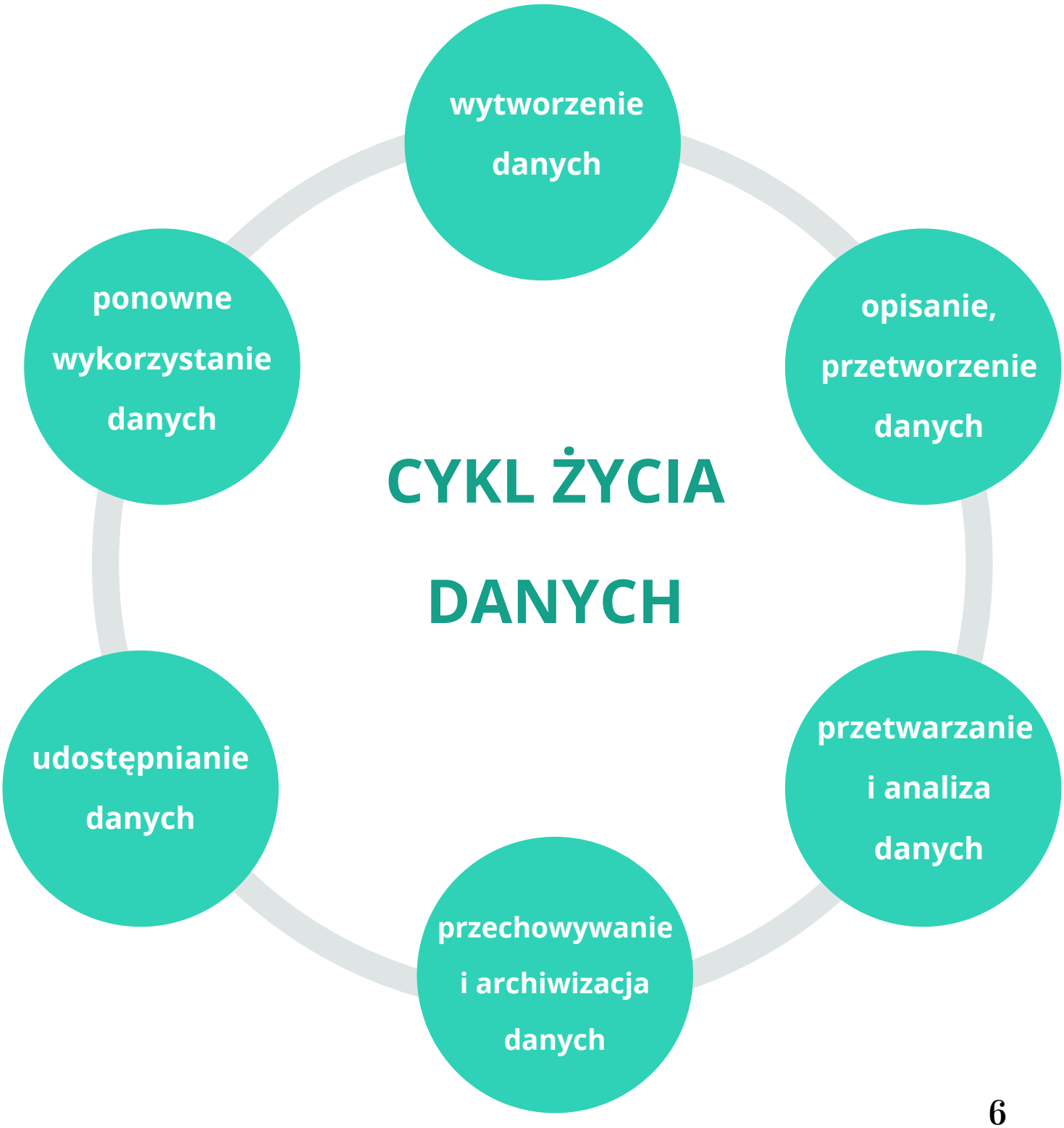
Dane badawcze można podzielić na:

- **surowe** – zebrane, ale nie przeanalizowane;
- **obserwacyjne** – przechwytywane w czasie rzeczywistym (np. odczyty czujników, dane telemetryczne, wyniki anonimowych ankiet, badania fokusowe), często unikalne, ponieważ nie można ich „odzyskać”;
- **eksperymentalne** – uzyskane ze sprzętu laboratoryjnego w kontrolowanych warunkach, powtarzalne, ale często bardzo kosztowne (np. sekwencje genów, spektroskopia, odczyty pola magnetycznego
- **dane symulacji** – zebrane podczas testów badających rzeczywiste lub teoretyczne systemy (np. modele klimatyczne, ekonomiczne, systemy inżynieryjne);
- **dane pochodne / skompilowane** – wyniki analiz danych, albo dane agregowane z różnych źródeł. Powtarzalne, ale ich pozyskanie może być bardzo kosztowne (bazy danych, teksty, modele 3D, dane bibliometryczne);
- **dane referencyjne** – poprawione lub organiczne zbiory danych, zwykle recenzowane, publikowane i selekcjonowane (dane GUS, struktury chemiczne, bazy danych z sekwencjami genów).

CYKL ŻYCIA DANYCH BADAWCZYCH

Dane często mają dłuższą żywotność niż projekt badawczy, który je tworzy. Naukowcy mogą kontynuować pracę nad danymi po zakończeniu finansowania, kolejne projekty mogą analizować lub uzupełniać te dane, które z kolei mogą być ponownie wykorzystywane przez innych badaczy.

Dobrze zorganizowane, udokumentowane, zarchiwizowane i udostępnione dane są nieocenione dla rozwoju nauki i przyspieszenia badań.



PLAN ZARZĄDZANIA DANYMI BADAWCZYMI

Dobrze zarządzane dane badawcze przynoszą wiele korzyści badaczowi, jego dyscyplinie i całemu społeczeństwu.

Planowanie powinno rozpocząć się już na etapie projektowania samych badań. Wczesne zaplanowanie zarządzania danymi badawczymi, w tym zasad ich udostępniania w przyszłości, jest kluczem do zapewnienia optymalnej obsługi danych badawczych przez cały czas trwania projektu badawczego i umożliwienia ponownego wykorzystania danych do przyszłych badań.

Zarządzanie danymi obejmuje wszystkie aspekty przetwarzania, organizowania, dokumentowania i ulepszania danych badawczych oraz gwarancję ich trwałości i udostępniania.

Planowanie najlepiej wykonać przy użyciu planu zarządzania danymi.

Plan zarządzania danymi ma charakter „żywego" dokumentu, który zmienia się w trakcie trwania projektu. We wszystkich planach zarządzania danymi badawczymi można znaleźć pewne wspólne elementy:

- zagadnienia związane z wytworzeniem / pozyskaniem danych,
- formaty plików i ich nazewnictwo,
- przechowywanie danych,
- dostęp do danych i ich ponowne wykorzystanie,
- aspekty prawne i etyczne,
- kwestie finansowe, infrastrukturalne i kompetencyjne (za ile?, gdzie?, kto za to odpowiada?).

ZALETY PLANU ZARZĄDZANIA DANYMI BADAWCZYMI

Plan zarządzania danymi ułatwia:

- pisanie publikacji przy użyciu danych zapisanych i konsekwentnie dokumentowanych w trakcie całego projektu,
- kontynuowanie pracy z danymi w razie sytuacji opuszczenia zespołu przez odpowiedzialnego za nie naukowca,
- udowodnienie wyników badań,
- wyselekcjonowanie danych do długoterminowej archiwizacji i do dalszego udostępniania
- komercjalizację wyników badań.

Plan zarządzania danymi zapobiega:

- utracie danych,
- kosztownemu powtarzaniu gromadzenia danych,
- przypadkowemu naruszeniu prywatności i przepisów etycznych.

Odpowiednie zarządzanie danymi badawczymi daje pewność, że dane są wiarygodne i kompletne oraz pozwala na budowanie przyszłych badań na solidnych podstawach.

Przejrzystość badań finansowanych ze środków publicznych jest ważna zarówno na poziomie indywidualnym, jak i instytucjonalnym.
Korzyści z posiadania planu ZDB dla pracowników naukowych:

- dodatkowe cytowania zarówno artykułów, jak i zestawów danych,
- widoczny wpływ poprzez dalsze komercyjne wykorzystanie danych,
- nieoczekiwane obserwacje naukowe wynikające z nowych technik i kombinacji zestawów danych.

CHECKLISTA PLANU BADAWCZEGO

Na liście znajdują się zagadnienia wspólne dla wszystkich planów zarządzania danymi badawczymi. Szczegółowość opisu i jego zakres zależą od prowadzonych badań i od wymagań grantodawcy, jednak zastanowienie się nad odpowiedziami na te pytania, pozwoli szybciej napisać plan zarządzania danymi badawczymi.

DANE PODSTAWOWE, OPIS PROJEKTU

Informacje podstawowe opisujące plan:

- Tytuł projektu
- Nazwisko i imię kierownika projektu / autora planu,
- Dane kontaktowe
- Numer projektu / grantu / ID
- Podsumowanie projektu opisujące cel zbierania danych
- Opis instytucjonalnej polityki zarządzania danymi badawczymi.

GROMADZENIE DANYCH

Do rozważenia:

- Jakie dane będą zbierane podczas badań?
- Jak będą gromadzone?
- Czy istnieją już dane, które możesz ponownie wykorzystać?
- Jakich standardów lub metodologii trzeba użyć, aby wytworzyć dane?
- Czy wybrane formaty i oprogramowania umożliwiają udostępnianie i długotrwałe przechowywanie danych?
- Jaka będzie strukturyzacja oraz nazewnictwo folderów i plików, w których przechowywane będą dane?
- Jakie procesy zostaną zastosowane, aby zapewnić odpowiednią jakości danych?

DOKUMENTACJA I METADANE

Do rozważenia:

- Jakie informacje są niezbędne, aby dane mogły być odczytane i zinterpretowane w przyszłości?
- Ile czasu i wysiłku będzie potrzeba na stworzenie dokumentacji pomocniczej i czy jest odpowiedni ich zasób (czasu i wysiłku).
- Jaka dokumentacja i metadane będą opisywać dane?
- W jaki sposób powstanie ta dokumentacja i metadane?
- Które standardy metadanych będą użyte i dlaczego te?

ZAGADNIENIA ETYCZNE I PRAWNE

Należy rozważyć wszystkie zagadnienia prawne i etyczne wiążące się z pozyskiwaniem danych w projekcie. Ważne są zwłaszcza ograniczenia udostępniania danych.

- Czy są wymagane zgody na udostępnianie i przechowywanie danych?
- Jak będzie chroniona tożsamość uczestników badań? (np. czy zostanie zastosowana anonimizacja?)
- Czy udostępnianie danych zostanie obłożone embargiem lub ograniczone (np. z powodu publikacji lub ubiegania się o patent)?
- Jaka licencja na wykorzystywanie danych zostanie wykorzystana?

PRZECHOWYWANIE DANYCH I TWORZENIE KOPII ZAPASOWYCH

Należy rozważyć:

- Gdzie będą przechowywane dane, jaki ma to wpływ na tworzenie kopii zapasowych, dostęp do danych i ich bezpieczeństwo?
- Czy jest miejsce na przechowywanie danych, czy potrzebne są fundusze na pokrycie kosztów przechowywania danych?
- Kto będzie odpowiedzialny za tworzenie backupów i odzyskiwanie danych?
- Jakie są zagrożenia dla bezpieczeństwa danych i jak nimi zarządzać?
- W jaki sposób będzie zapewniony dostęp do danych dla wszystkich współpracowników?

SELEKCJA DANYCH I ICH OCHRONA

Nalezy określić, które dane będą długotrwale przechowywane i chronione.

- Należy wybrać najlepszy sposób przechowywania danych (wybór repozytorium).
- Które dane należy zachować, a które zniszczyć z powodów wynikających z umów i regulacji prawnych?
- Jakie są przewidziane inne zastosowania badawcze dla danych?
- Które dane powinny zostać zachowane i potencjalnie udostępnione?
- Jaki jest plan długoterminowego przechowywania bazy danych?
- Jaki jest koszt przygotowania, przechowywania i udostępniania twoich danych?

UDOSTĘPNIANIE DANYCH

Które dane będą udostępniane i w jaki spoób? Wybór metody zależy od wielu czynników, takich jak: typ, rozmiar, złożoność i wrażliwość danych. Nalezy rozważyć:

- W jaki sposób będą cytowane dane?
- Z kim będą współdzielone dane i na jakich warunkach?
- Kiedy będzie otwarty dostęp do danych?
- Czy wymagane są jakieś ograniczenia dotyczące udostępniania danych?
- Jakie działania będa podejmowane, aby pokonać lub zminimalizować ograniczenia w dostępie?
- Jak potencjalni użytkownicy dowiedzą się o danych?

OBOWIĄZKI I ZASOBY DANYCH

Należy przypisać role i obowiązki dla wszystkich działań związanych z zarządzaniem danymi. Koszty zazwyczaj można wpisać do wniosku grantowego, ale muszą być jasno określone i uzasadnione.Do rozważenia, co jest potrzebne, aby zrealizować plan zarządzania danymi.

- Kto jest odpowiedzialny za zrealizowanie planu zarządzania oraz za jego sprawdzenie i poprawienie?
- Jak zostaną rozdzielone obowiązki między partnerami w projekcie badawczym?
- Jakie zasoby są potrzebne, aby zrealizować plan?
- Czy wymagana jest dodatkowa wiedza specjalistyczna lub sprzęt?

ORGANIZACJA DANYCH BADAWCZYCH

Dane badawcze powinny być udostępniane w formie tak zwanych datasetów, czyli w uporządkowanej i dobrze opisanej strukturze.

Dobre opisanie i uporządkowanie danych ułatwi ich przyszłe wyszukiwanie, odczytanie i wykorzystanie. Sposób, w jaki dane są zapisywane i przedstawiane ma też duży wpływ na możliwe sposoby przetwarzania i analizy danych.

Aby dane były przydatne powinny być uzupełnione o następujące informacje kontekstowe:

- w jakim celu dane zostały przechwycone / wygenerowane, przez kogo i kiedy,
- informacje niezbędne do interpretacji danych (np. warunki eksperymentalne, próbkowanie statystyczne, informacje o kalibracji),
- prawa i obowiązki związane z danymi, w tym licencjonowanie (jeśli dane są udostępniane) lub warunki dostępu (jeśli dostęp jest ograniczony).

Informacje o danych powinny być regularnie aktualizowane oraz powinny być zapisane w pliku łatwym do odczytu maszynowego.

SELEKCJA DANYCH

Podczas badań wytwarzane są duże ilości danych różnego rodzaju. Pierwszym krokiem do opracowania planu zarządzania danymi powinna być selekcja danych, czyli analiza, które dane należy zachować, w jakim formacie oraz kto powinien nimi zarządzać, a kto mieć tylko dostęp.

Wybierając dane do archiwizacji warto wziąć pod uwagę:

- **Wymagania prawne zobowiązujące nas do archiwizacji danych** – czy zawartość zasobów odpowiada kompetencji ośrodka i wszelkim priorytetom określonym w obecnej strategii instytucji badawczej lub podmiotu finansującego, w tym wszelkim wymogom prawnym dotyczącym zatrzymywania danych poza ich bezpośrednim wykorzystaniem.

- **Wartość naukową lub historyczną danych –** czy dane są istotne z naukowego, społecznego lub kulturowego punktu widzenia?

- **Wyjątkowość –** czy dane duplikują się z innymi istniejącymi zbiorami danych? Czy istnieje ryzyko utraty danych, jeśli nie zostaną zarchiwizowane?

- **Możliwość wykorzystania** – czy formaty danych są od strony technicznej dobrze dobrane? Czy kwestie praw własności intelektualnej są wyjaśnione?

- **Możliwość replikacji** – czy można takie dane ponownie zebrać (wysokie koszty, jednorazowe wydarzenie)?

- **Kwestie ekonomiczne** – czy koszty zarządzania danymi i ich przechowywania są uzasadnione w świetle potencjalnego przyszłego wykorzystania danych?

- **Pełna dokumentacja –** czy dokumentacja jest poprawna i kompletna, w tym metadane dotyczące pochodzenia zasobu oraz kontekst jego tworzenia i użytkowania?

METADANE

Metadane, czyli dane o danych, to istotny element zbioru danych. Zawierają informacje o formie i treści zasobów, dzięki czemu umożliwiają ich wyszukiwanie i identyfikację oraz zarządzanie nimi.

Standardy metadanych służą usystematyzowaniu sposobu opisu danych. Posiadają stałą strukturę opisu o wyraźnie zdefiniowanych polach, dzięki czemu opis jest zawsze zrozumiały zarówno dla ludzi jak i programów komputerowych.

Nie ma jednego standardu dla metadanych. Wyszczególnić można standardy ogólne, dziedzinowe i instytucjonalne. Ogólne standardy metadanych to ***Dublin Core*** i ***Data Cite***, ***Data Documentation Initiative (DDI)***. Są one uniwersalne dziedzinowo i powszechnie stosowane.

Powstało wiele inicjatyw mających na celu sformalizowanie specyfikacji metadanych, aby umożliwić łatwe ponowne wykorzystanie danych. Przykładami takich inicjatyw są: ***Research Data Alliance (RDA)***, ***OpenAire*** i ***Metadata 2020***.

Niektóre instytucje czy dyscypliny opracowały też własne standardy, np.:

- DC (nauki o życiu),
- EML (ekologia),
- SDMX (ECB, EUROSTAT, IMF, OECD, UN),
- INSPIRE ISO 19139 (nauki o ziemi),
- Project Open Data Metadata Schema v1.1 (Agencje federalne USA),
- CDWA (dyscypliny humanistyczne).

Wszystkie dane badawcze muszą być udostępnione wraz z ich metadanymi. Metadane powinny być zawsze dostępne, nawet jeśli dane, które opisują nie są już osiągalne.

Metadane dzielimy na:

- **Metadane opisowe** – dostarczają informacji niezbędnych do odszukania czy też identyfikacji zbioru danych. Mogą zawierać elementy, takie jak tytuł, streszczenie, autor i słowa kluczowe.

- **Metadane strukturalne** – opisują relacje i zależności pomiędzy poszczególnymi zbiorami oraz elementami tych zbiorów np. w celu ułatwienia nawigacji.

- **Metadane administracyjne** – zawierają informacje pomocne w zarządzaniu danym zasobem. Zawierają informacje takie jak sposób i datę jego utworzenia, typ pliku i informacje dotyczące dostępu. Istnieje kilka podzbiorów danych administracyjnych. Dwa z nich są często wymieniane jako oddzielne typy metadanych:
  - metadane zarządzania prawami, które dotyczą praw własności intelektualnej,
  - metadane konserwacji, które zawierają informacje potrzebne do archiwizacji i utrzymania zasobu.

FORMAT PLIKÓW

Dane badawcze istnieją w wielu różnych formach: tekstowych, numerycznych, w formie obrazów, nagrań audiowizualnych. Aby można było z nich w przyszłości bezproblemowo korzystać, warto zapisywać danew ogólnodostępnym formatach plików,łatwym do odczytania i interpretowania. Korzystanie ze standardowych i wymiennych lub otwartych formatów danych bezstratnych zapewnia długoterminową użyteczność danych. Należy wybrać formaty:

- bez kompresji,
- nie wymagające komercyjnego oprogramowania,
- otwarte, z dostępną dokumentacją,
- wykorzystujące standardowe kodowanie (ASCII, Unicode).

	FORMAT PREFEROWANY	FORMAT AKCEPTOWALNY
DANE LICZBOWE	.csv, .tsv, .spss, .por	.xls, .sav, .dta, .mdb/.accdb
DANE TEKSTOWE	.odt, .ods	.doc, .docx, .pdf, .xml, .htm, .html,. rtf, .xlsx, .epub
DANE GEOPRZESTRZENNE	.shp, .shx, .dbf, .sbn, .sbx, .prj, .xml	PostGIS, tif, .tfw, .fdg, .adf, .dat, .nit
PLIKI AUDIO	.wav, .aif, .aiff, .flac	.mp3, .m4p, .m4a, .mid, .midi, .ogg
PLIKI WIDEO	.avi	.mov, .wmv,. mpg
DANE OBRAZU	.tiff, .jpeg2000, .png, .svg	.gif, .jpg, .ai, .cgm
PREZENTACJE	.pdf, .odp	.pptx

Niektóre repozytoria umożliwiają deponowanie danych w dwóch wersjach:

- w formacie przeznaczonym do długotrwałej archiwizacji,
- w formacie najpowszechniej wykorzystywanym w danym środowisku.

KOSZTY

Działania związane z zarządzaniem danymi i dzieleniem się nimi generują koszty. Dobrą praktyką jest uwzględnienie tych kosztów w planie zarządzania danymi badawczymi. Można wyróżnić dwa podejścia:

Podejście 1 – wyliczenie kosztów obejmuje wszystkie działania związane z wytwarzaniem, przechowywaniem i udostępnianiem danych, czyli wszystkie działania i zasoby związane z danymi w całym cyklu życia danych - od tworzenia danych, przez przetwarzanie, analizy i przechowywanie, po udostępnianie i długoterminowe przechowywanie.

Podejście 2 – koszty obejmują wyłącznie te elementy, które byłyby potrzebne do zachowania i udostępnienia danych badawczych poza głównym zespołem badawczym. Zasoby te mogą obejmować: osoby, sprzęt, infrastrukturę i narzędzia do zarządzania, dokumentowania, organizowania, przechowywania i zapewniania dostępu do danych.

Oba mogą być wykorzystane w planie zarządzania danymi lub mogą stanowić podstawę wniosku o finansowanie.

ROLE I OBOWIĄZKI

Zarządzanie danymi powinno leżeć w gestii nie tylko badacza, który utworzył lub zgromadził dane. Część obowiązku spoczywa na tych podmiotach, które uczetniczyły w procesie badawczym na różnych jego etapach. W planie zarządzania danymi powinno znaleźć się wyraźne rozpisanie ról i obowiązków między partnerami projektu.

Osoby zaangażowane w zarządzanie danymi i ich udostępnianie to:

- dyrektor projektu planujący i nadzorujący badania,
- personel badawczy projektujący badania, gromadzący, przetwarzający i analizujący dane,
- personel laboratoryjny lub techniczny tworzący metadane i dokumentację,
- projektant bazy danych,
- pracownicy instytucjonalnych służb IT, świadczący usługi przechowywania danych, bezpieczeństwa i tworzenia kopii zapasowych,
- personel zarządzający badaniami i finansowaniem badań oraz administrujący nimi, zapewniający przegląd etyczny i ocenę praw własności intelektualnej,
- zewnętrzne centra danych lub archiwa internetowe, które ułatwiają udostępnianie danych.

NARZĘDZIA POMOCNE W TWORZENIU PLANU ZARZĄDZANIA DANYMI

DMPtool (US) – narzędzie przygotowujące szablony DMP dostosowane do wymagań amerykańskich grantodawców.

DMPonline (UK) – narzędzie bardzo podobne do DMPtool zawierające jednak bazę instytucji finansującej naukę z Wielkiej Brytanii.

DCC Data Management Plan Content Checklist – lista kontrolna zawartości planu zarządzania danymi, pozwala szybko określić, jakich informacji może brakować w przygotowywanym DMP.

The Data Curation Center – serwis brytyjskiej instytucji specjalizującej się w zarządzaniu danymi badawczymi. Udostępnia między innymi: gotowe plany zarządzania danymi, przewodniki, wytyczne, informacje na temat metadanych.

Listę przydatnych narzędzi znajdziesz na: buw.uw.edu.pl

ASPEKTY PRAWNE

Niektóre z gromadzonych danych badawczych to dane wrażliwe lub dane chronione prawem autorskim.

Nie oznacza to, że takie dane w ogóle nie mogą być udostępniane. Wykorzystując odpowiednie mechanizmy, dane zawierające informacje wrażliwe (np. informacje zdrowotne, orientację seksualną, pochodzenie etniczne, religię itp.) można udostępnić etycznie i legalnie, jeśli naukowcy zastosują strategię świadomej zgody, anonimizacji i kontroli dostępu do danych.

Zgoda na gromadzenie, przetwarzanie i wykorzystywanie danych

Przepisy dotyczące ochrony danych oraz ogólne rozporządzenie o ochronie danych (RODO) wymusza otrzymanie od uczestników badania zgody na gromadzenie i przyszłe ponowne wykorzystanie danych przez innych badaczy. Uczestnicy muszą być poinformowani, w jaki sposób dane badawcze będą przechowywane i wykorzystywane w perspektywie długoterminowej oraz w jaki sposób zostanie zachowana poufność.

Kontrola dostępu

Dane wrażliwe i poufne można zabezpieczyć, regulując lub ograniczając do nich dostęp i ich wykorzystanie. Kontrola dostępu powinna zawsze być proporcjonalna do rodzaju danych i poziomu poufności.

Anonimizacja

Anonimizacja danych to proces polegający na przekształceniu danych osobowych w sposób uniemożliwiający przyporządkowanie poszczególnych informacji do określonej lub możliwej do zidentyfikowania osoby. Formą anonimizacji może być: używanie pseudonimów zamiast nazwisk, usunięcie kluczowych zmiennych lub rozmycie danych obrazu lub wideo. Anonimizacja to proces nieodwracalny.

Procesem odwracalnym jest ***pseudonimizacja***, czyli takie przetworzenie danych, by nie można ich było przypisać osobie, której te dane dotyczą, bez użycia dodatkowych informacji. Takie dodatkowe informacje należy przechowywać osobno i zabezpieczyć środkami technicznymi i organizacyjnymi, uniemożliwiającymi ich przypisanie zidentyfikowanej lub możliwej do zidentyfikowania osobie fizycznej. Formą pseudonimizacji jest zamiana danych (np. imienia i nazwiska) na ciąg liter lub cyfr, które można rozszyfrować wyłącznie na podstawie przechowywanego oddzielnie klucza.

LICENCJONOWANIE DANYCH

Dane badawcze, podobnie jak inne aspekty działalności naukowej, podlegają przepisom prawa. Niektóre dane gromadzone w ramach projektu badawczego podlegają tak samo prawu własności intelektualnej, jak dzieła literackie lub artystyczne. Dane medyczne podlegają natomiast prawu do prywatności i prawu tajemnicy lekarskiej.

Przygotowując plan danych badawczych należy nie tylko oznaczyć prawa, jakim podlegają dane, ale również uwzględnić zasady, na jakich będzie można z nich korzystać. Można stworzyć własne „zasady" korzystania z danych lub, co jest praktyczniejsze, skorzystać z opracowanych już licencji.

Organizacje zajmujące się popularyzacją otwartej nauki rekomendują wybór jednego z poniższych rozwiązań:

Licencje Creative Commons – to najpopularniejsze licencje, wykorzystywane szeroko w nauce i edukacji. Biorąc pod uwagę zalecenia otwartości danych poleca się stosowanie jednej z tych licencji:

- uznanie autorstwa **CC BY**,
- uznanie autorstwa-na tych samych warunkach **CC BY-SA**,
- przekazanie do Domeny Publicznej **CC0**.

Open Data Commons – (www.opendatacommons.org) to projekt, w którym stworzono zestaw narzędzi prawnych (licencji) wspomagających udostępnianie i korzystanie z otwartych danych:

- **Public Domain Dedication and License (PDDL)** – domena publiczna dla baz danych. Zakłada nieograniczoną możliwość pobierania, udostępniania i modyfikowania baz danych.

- **Open Data Commons Attribution License (ODC-By)** – licencja, w której jedynym warunkiem kopiowania i modyfikowania bazy danych jest uznanie autorstwa.

- **Open Data Commons Open Database License (ODC – ODbL)** – otwarta licencja zezwalająca na kopiowanie, przetwarzanie oraz rozpowszechnianie bazy danych pod warunkiem uznania jej autorstwa oraz upowszechniania wyników na takich samych warunkach.

- **Domena publiczna**.

UDOSTĘPNIANIE DANYCH

Udostępnianie danych sprzyja przejrzystości badań, przyspiesza dokonywanie kolejnych odkryć naukowych oraz ułatwia popularyzację wyników badań naukowych.

Otwieranie danych badawczych to jedna z praktyk otwartej nauki. Dostęp do wyników badań to kluczowa kwestia dla jakości i przejrzystości procesu badawczego. Dane uzyskane w badaniach finansowanych ze środków publicznych muszą być powszechnie dostępne.

OTWIERANIE DANYCH BADAWCZYCH

Otwarte dane badawcze to takie, które są dostępne w postaci cyfrowej w domenie publicznej, upowszechniane bez żadnych ograniczeń, przez co mogą być swobodnie używane, rozpowszechniane i przetwarzane przez kogokolwiek, gdziekolwiek w dowolnym celu.

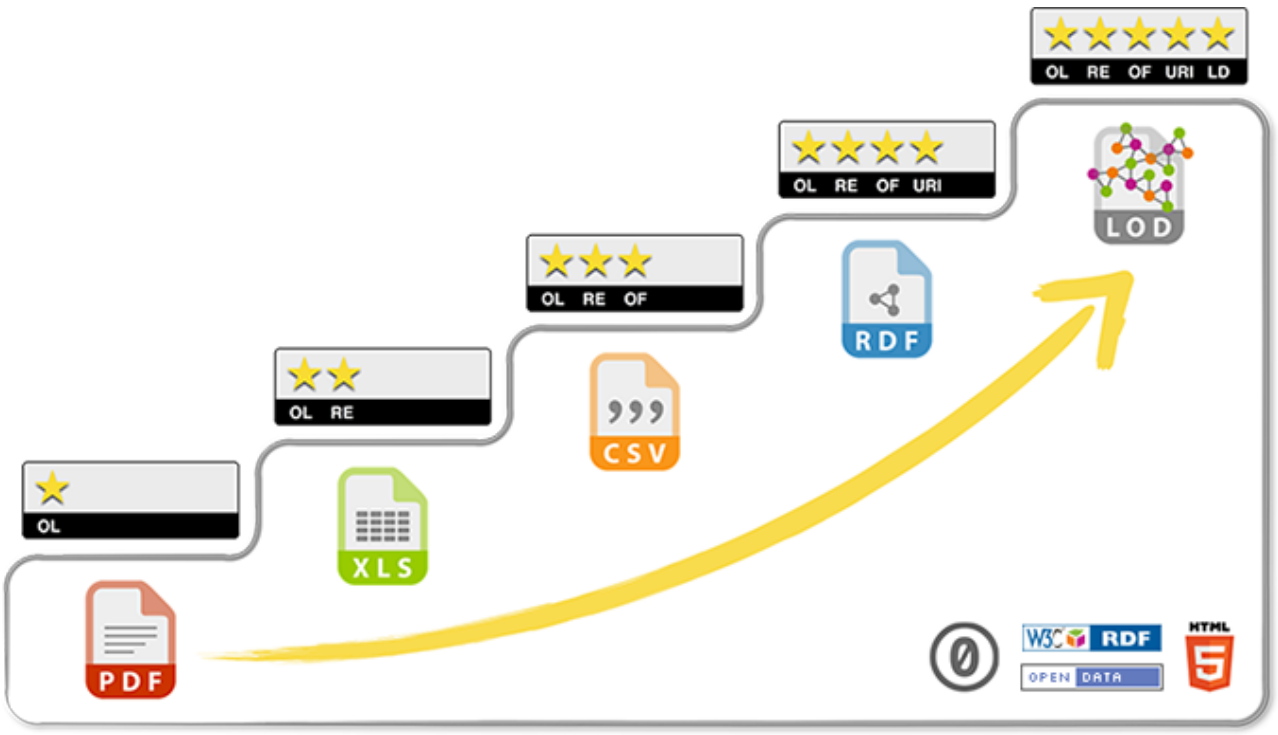
Otwartość danych oznacza nie tylko otwarty do nich dostęp, ale przede wszystkim swobodę ponownego ich wykorzystania.

PO CO UDOSTĘPNIAĆ DANE BADAWCZE?

- **Otwarte dane można wykorzystywać do prowadzenia nowych badań, a także łączyć je ze sobą, tworząc nowe zestawienia.**
- Udostępnienie danych umożliwia ich ponowną analizę i zachęca do nowych interpretacji.
- Udostępnienie danych ułatwia sprawdzenie, czy opublikowane już prace naukowe opierają się na powtarzalnych wynikach.

Tim Bernes'a Lee, brytyjski fizyk i programista, opracował pięciopoziomową skalę otwartości danych (tzw. „Five Stars Open Data"):

- **Poziom pierwszy** (*) – udostępnienie danych na stronach WWW w dowolnym ogólnodostępnym formacie (np. PDF-skan, HTML, JPG).
- **Poziom drugi** (**) – udostępnianie danych w postaci dowolnego pliku tabelarycznego (np. XLS).
- **Poziom trzeci** (***) – udostępnianie danych w postaci dowolnego pliku tabelarycznego, ale o otwartej strukturze, niezwiązanej z producentem konkretnego oprogramowania (np. CSV, XML).
- **Poziom czwarty** (****) – udostępnianie danych z wykorzystaniem odnośników URL i wykorzystaniem technologii RDF opisującej dane.
- **Poziom piąty** (*****) – udostępnienie danych w połączeniu z innymi danymi z aktualnych, otwartych zbiorów danych by zapewnić im kontekst.



Zasady FAIR Data to zbiór wytycznych dla opublikowanego zbioru danych, których spełnienie umożliwia ponowne wykorzystania danych zarówno przez ludzi, jak i przez maszyny.

F FINDABLE

Łatwe do znalezienia i wyszukania

Pierwszym krokiem w (ponownym) korzystaniu z danych jest ich znalezienie. Metadane i dane powinny być łatwe do odnalezienia zarówno dla ludzi, jak i dla komputerów. Metadane do odczytu maszynowego są niezbędne do automatycznego wykrywania zestawów danych i usług, dlatego jest to niezbędny element procesu FAIRification.

- F1. Zbiory danych i metadane mają unikatowy na całym świecie i trwały identyfikator (np. DOI)
- F2. Dane są opisane bogatymi metadanymi (zdefiniowanymi przez R1 poniżej)
- F3. Metadane w sposób jasny i wyraźny przedstawiają identyfikator opisywanych danych.
- F4. Zbiory danych i metadane są rejestrowane lub indeksowane w ogólnodostępnych bazach danych umożliwiających ich przeszukiwanie.

A ACCESIBLE

Dostępne dla wszystkich

Po znalezieniu wymaganych danych użytkownik musi zostać poinformowany o sposobie dostępu, w tym o ewentualnym uwierzytelnieniu i autoryzacji.

- A1. Zbiory danych i metadane można odzyskać po ich identyfikatorze przy użyciu standardowego protokołu komunikacyjnego.
- A1.1. Protokół jest otwarty, bezpłatny i uniwersalny do wdrożenia.
- A1.2. Protokół umożliwia w razie potrzeby procedurę uwierzytelnienia i autoryzacji.
- A2. Metadane są dostępne, nawet jeśli dane nie są już dostępne.

I INTEROPERABLE

Współpracujące z innymi danymi

Dane zwykle muszą być zintegrowane z innymi danymi. Ponadto dane muszą współpracować z aplikacjami lub przepływami pracy w celu analizy, przechowywania i przetwarzania.

- I1. Zbiory danych i metadane opisane są za pomocą formalnego, dostępnego, współdzielonego i powszechnie stosowanego języka.
- I2. Zbiory danych i metadane wykorzystują słowniki zgodne z zasadami FAIR.
- I3. Zbiory danych i metadane je opisujące zawierają odnośniki do
- innych powiązanych z nimi zbiorów.

R REUSABLE

Możliwe do ponownego wykorzystania

Ostatecznym celem FAIR jest optymalizacja ponownego wykorzystania danych. Aby to osiągnąć, metadane i dane powinny być dobrze opisane, aby można je było replikować i / lub łączyć w różnych zestawieniach.

- R1. Zbiory danych i metadane są bogato opisane za pomocą wielu dokładnych i odpowiednich atrybutów.
- R1.1. Zbiory danych posiadają jasną i dostępną licencję na wykorzystanie danych.
- R1.2. Zbiory danych i metadane mają wyraźnie oznaczone autorstwo i pochodzenie danych.
- R1.3. Zbiory danych i metadane są przygotowane zgodnie z przyjętymi standardami charakterystycznymi dla danej dyscypliny oraz rodzaju danych.

REPOZYTORIA DANYCH BADAWCZYCH

Repozytorium danych i archiwa zapewniają bezpieczne, długoterminowe przechowywanie. Umożliwiają kontrolowany dostęp do danych wrażliwych. Ponadto repozytoria udostępniają też często informacje statystyczne o tym, jak często dane były pobierane i oglądane.

Archiwa biorą odpowiedzialność za przetwarzanie zapytań o ponowne wykorzystanie danych, licencjonowanie, rozpowszechnianie i promocję danych w imieniu właściciela danych. Archiwa odpowiadają również za bezpieczne długoterminowe przechowywanie, chronią dane przed utratą, pogorszeniem lub ich nieodwracalnym uszkodzeniem.

Można wyróżnić:

- repozytoria dziedzinowe,
- repozytorium instytucjonalne,
- repozytoria ogólnego przeznaczenia.

Szukając odpowiedniego repozytorium warto:

- dokładnie zapoznać się z warunkami korzystania z serwisu i sprawdzić czy spełnia on nasze wymagania;
- dowiedzieć się gdzie i na jakich zasadach będą przechowywane nasze dane oraz w jaki sposób będą zabezpieczone;
- upewnić się, że dane repozytorium zapewnia przypisanie naszym zbiorom identyfikatora DOI;
- sprawdzić czy inni naukowcy z naszej dyscypliny korzystają z danego repozytorium;
- dowiedzieć się, czy repozytorium wspiera używany w naszej dyscyplinie standard metadanych;
- mieć na uwadze, że niektóre repozytoria mogą pobierać opłatę za archiwizację danych – tzw. Data Processing Charge;
- upewnić się, że zasoby wybranego przez nas repozytorium znajdują się w bazach indeksujących repozytoria danych badawczych. Przykładami takich baz są Data Citation Index, Mendeley Data czy Google Dataset Search.

Listę repozytoriów znajdziesz na: buw.uw.edu.pl

SAMOZACHOWYWANIE I ROZPOWSZECHNIANIE

Dane można udostępniać za pośrednictwem stron internetowych projektu lub nieformalnego udostępniania *peer-to-peer*.

Przechowywanie danych na stronie projektu, a nie w repozytoriach, może jednak wpływać na mniejszą trwałość danych i ich długoterminową ochronę. Może to być również kosztowne w zarządzaniu, a także trudne do kontrolowania, kto i jak korzysta z danych.

Nieformalne udostępnianie peer-to-peer umożliwia szybkie udostępnianie. Utrudnia to jednak odbiorcom ustalenie, które dane można uzyskać, z kim się skontaktować, a zarządzanie dostępem do danych staje się dużym obciążeniem i nie zapewnia długoterminowej dostępności danych.

CZASOPISMA PUBLIKUJĄCE DANE (DATA JOURNALS)

Coraz więcej wydawców ma w swojej ofercie czasopisma typu data journals publikujące zestawy danych badawczych. Są to czasopisma recenzowane działające na wzór tradycyjnych czasopism publikujących artykuły. Dane są deponowane w repozytoriach, a niektóre czasopisma dopuszczają możliwość dołączania danych w postaci Supplementary Material. Publikowanie danych w data journals to dobra praktyka promująca badania i uzupełniająca standardowe deponowanie danych w repozytoriach.

DANE BADAWCZE W PROJEKTACH KRAJOWYCH I EUROPEJSKICH

Instytucje i programy finansujące badania naukowe coraz częściej wymagają od naukowców przedstawienia planu danych badawczych na etapie składania i oceny wniosków grantowych.

Wiele z instytucji idzie dalej w swoich założeniach i nie tylko wymaga samego planu zarządzania danymi, ale również udostępniania danych badawczych w otwartym dostępie. Takie działania mają na celu zwiększenie widoczności wyników badań, a także ułatwiają i przyśpieszają komunikację naukową oraz współpracę pomiędzy badaczami.

Przyjęta 20 czerwca 2019 przez Parlament Europejski Radę (UE) **Dyrektywa w sprawie otwartych danych i ponownego wykorzystywania informacji sektora** publicznego (2019/1024) włącza dane badawcze do danych sektora publicznego i nakłada na kraje członkowskie obowiązek zapewniena możliwość ponownego ich wykorzystania.

Dyrektywa zobowiązuje kraje UE do wdrożenia regulacji prawnych i działań, których celem ma być gwarancja powszechnego dostępu do danych na możliwie najwcześniejszym etapie ich rozpowszechniania. Politykę otwartości powinny przyjąć wszystkie organizacje prowadzące i finansujące badania naukowe ze środków publicznych.

Obowiązek udostępnienia danych dotyczy tych dokumentów, które powstały dzięki finansowaniu ze środków publicznych i są już dostępne w repozytoriach instytucjonalnych lub dziedzinowych. Z zobowiązku udostępniania zwolnione są prace publikowane w czasopismach naukowych ze względu na dodatkowe wyzwania związane z zarządzaniem prawami.

Nowa dyrektywa ma wejść w życie w czerwcu 2021 r.

HORIZON 2020

Program HORIZON 2020 to Program Ramowy Unii Europejskiej w zakresie badań naukowych i innowacji. Jednym z jego założeń jest zapewnienie otwartego dostępu do wyników badań prowadzonych w ramach programu. Od roku 2017 program HORIZON 2020 rozszerzono o dane badawcze.

Pilotaż Otwartych Danych Badawczych (**Open Research Data Pilot**) wymaga udostępnienia w domenie publicznej danych badawczych z badań finansowanych w ramach programu Horyzont 2020.

Pilotaż Otwartych Danych obejmuje dwa rodzaje danych:
1) dane (...) niezbędne do weryfikacji wyników prezentowanych w publikacjach naukowych należy udostępniać tak szybko, jak to możliwe;
2) inne dane (...) wymienione w planie zarządzania danymi należy udostępniać zgodnie z ustalonymi w planie terminami.
(...) Projekty objęte pilotażem są zobowiązane do deponowania opisanych powyżej danych badawczych, najlepiej w repozytoriach danych badawczych."

Dane badawcze powinny być tak otwarte, jak to możliwe i na tyle zamknięte, na ile to jest konieczne. Udostępnienie powinno nastąpić jak najszybciej, ale nie później niż 6 miesięcy po publikacji wyników. Zachowano również możliwość rezygnacji (opt out) z udostępniania danych w uzasadnionych sytuacjach.

Z udostępniania danych wyłączone są:
- komercyjne lub przemysłowe wykorzystanie danych,
- wymogi poufności związane z bezpieczeństwem,
- ochrona danych osobowych,
- niemożliwość osiągnięcia głównego celu,
- brak danych.

Zgodnie z wytycznymi zawartymi w **Open Research Data Pilot** uczestnicy są zobowiązani do:
- przygotowania i aktualizowania **Planu zarządzania danymi**,
- zdeponowania danych w repozytoriach danych badawczych,
- określenia zasad swobodnego wykorzystywania danych (w tym licencje CC-BY lub oświadczenia CC0),
- określenia, jakich narzędzi należy użyć w celu weryfikacji danych surowych (lub dostarczenie takich narzędzi).

NARODOWE CENTRUM NAUKI

Narodowe Centrum Nauki jest członkiem Science Europe, organizacji zrzeszającej europejskie instytucje naukowe finansujące lub prowadzące badania, które działają wspólnie m.in. na rzecz otwartej nauki.

W 2018 NCN było jednym z sygnotariuszy tzw. **cOAlition S**, czyli porozumienia instytucji finansujących badania naukowe, chcących wprowadzić od 2021 roku otwarty dostęp do wyników badań.

Zgodnie z wymogami NCN plan danych badawczych jest obowiązkowym elementem wniosku grantowego. Podlega eksperckiej ocenie jako merytoryczna część raportu końcowego. Jeśli plan zarządzania danymi będzie niekompletny, raport zostanie odesłany do korekty.

Centrum zaleca uaktualnianie planu zarządzania danymi w trakcie trwania projektu - w raporcie końcowym należy opisać stan faktyczny dot. danych w projekcie na dzień zakończenia projektu.

W ocenie PZD brane pod uwagę będą:

- opis danych, pozyskiwanie i ponowne ich wykorzystanie,
- dokumentacja i jakość danych,
- przechowywanie i kopie zapasowe,
- wymogi prawne, kodeksy postępowania,
- udostępnianie i długotrwałe przechowywanie,
- zarządzanie danymi.

Według wytycznych NCN otwarty dostęp powinien być zapewniony przede wszystkim do tych danych, które stanowią podstawę opublikowanych wyników i powinne być udostępniane w momencie ukazania się publikacji. Rekomendowany czas przechowywania danych to 10 lat.

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.

SŁOWNICZEK

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 w trakcie 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

BIBLIOGRAFIA

Skorzystaj z bazy wiedzy na temat otwarej nauki i otwartych danych badawczych na www.buw.edu.pl

Akty urzędowe

- Deklaracja sorbońska dotycząca praw do danych badawczych
- Deklaracja Sorbońska w sprawie danych badawczych
- H2020 Programme Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon
- Pilotaż otwartych danych badawczych w programie Horyzont 2020
- Pismo Dyrektora NCN w sprawie zarządzania danymi naukowymi w projektach
- Przewodnik dot. ujednoliconych europejskich praktyk związanych z zarządzaniem danymi badawczymi
- Standard techniczny otwartych danych, Ministerstwo Cyfryzacji
- Wytyczne dla wnioskodawców do uzupełnienia PLANU ZARZĄDZANIA DANYMI w projekcie badawczym.

Artykuły

- Dobre praktyki publikowania danych badawczych, Anna Małgorzata Kamińska, Biuletyn EBIB, nr 7 (177)/2017.
- Otwarte dane badawcze w warsztacie pracy naukowców, Małgorzata Rożniakowska-Kłosińska, Biuletyn EBIB, nr 6 (183)/2018.
- Otwieranie małych danych badawczych, Marta Hoffman-Sommer, Forum Akademickie 07-08/2016.
- The risks of not sharing data are greater than the costs, Paul Ayris.

Książki

- Exploring research data management, Andrew M. Cox and Eddy Verbaan, 2018
- Jak korzystać z zasobów w repozytoriach danych.
- Plan zarządzania danymi badawczymi, Repozytorium KUL.
- Plan zarządzania danymi badawczymi , Poradnik Wydawnictwa KUL.
- Prawne aspekty otwierania danych badawczych
- SCIENCE EUROPE - przewodnik nt. danych badawczych
- Selekcja i przygotowanie danych badawczych do udostępniania
- Udostępnianie danych badawczych – zagadnienia prawne (N. Rycko)
- Zarządzanie danymi badawczymi (N. Gruenpeter)

BIBLIOGRAFIA

Raporty

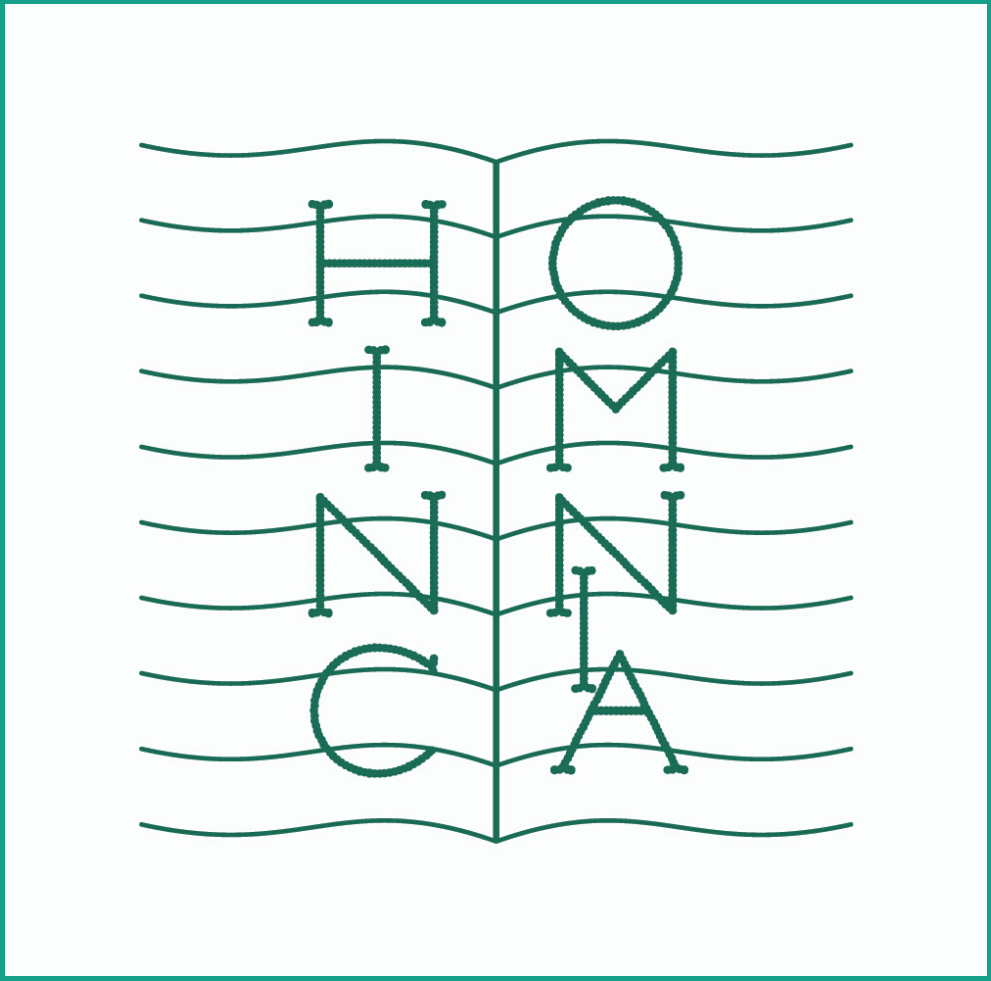
- Cost of not having FAIR research data.

Szkolenia

- Data Management Plan (DMP) w bibliotece naukowej. Nowe zadania i narzędzia, Edyta Rogowska, Tomasz Nowocień.
- MANTRA.
- Otwarte dane badawcze w humanistyce, Natalia Gruenpeter, Michał Starczewski.
- Otwarty kurs DMP na platformie Moodle The London School of Economics and Political Science.
- Research Data Management and Sharing.
- Udostępnianie danych badawczych– zagadnienia prawne, Nikodem Rycko.
- Warsztaty z zarządzania danymi badawczymi, Lublin, Natalia Gruenpeter.
- Warsztaty z zarządzania danymi badawczymi, Natalia Gruenpeter, Łódź 11.06.2019
- Wykład Bożeny Bednarek-Michalskiej z Biblioteki UMK
- Wykład dr. Tomasza Miksy na temat danych badawczych i nowoczesnych, maszynowo przetwarzalnych DMP (Institute of Software Technology and Interactive Systems, Vienna University of Technology)
- Zarządzanie danymi badawczymi Platforma Otwartej Nauki, Marta
- Hoffman-Sommer
- Zarządzanie danymi badawczymi, Bożena Bednarek-Michalska
- Zarządzanie metadanymi - wprowadzenie. Open Data Support.

Strony

- „Dziedzinowe Repozytoria Otwartych Danych Badawczych”
- Data management - H2020 Online Manual
- Data Management Plan, Pomorski Uniwersytet Medyczny
- Plan Zarządzania Danymi, UMK
- Repozytoria rekomendowane przez NATURE dla nauk ścisłych i przyrodniczych
- Research Data Management Working Group, LIBER
- Zalecane standardy metadanych w poszczególnych dyscyplinach



Warszawa 2020