Dane badawcze w pigułce

Poradnik

Polska Platforma Medyczna

Dane badawcze w pigułce. Poradnik
Opracowanie: zespół PPM

Data powstania: 27.04.2020 r.
Data ostatniej aktualizacji: 10.07.2020 r.

Poradnik powstał w ramach projektu: Polska Platforma Medyczna
http://www.ppm.edu.pl

Projekt „Polska Platforma Medyczna: portal zarządzania wiedzą i potencjałem badawczym" realizowany jest w oparciu o umowę nr POPC.02.03.01-00-0008/17-00 Program Operacyjny Polska Cyfrowa na lata 2014-2020, Oś Priorytetowa nr 2 „E-administracja i otwarty rząd", Działanie 2.3 „Cyfrowa dostępność i użyteczność informacji sektora publicznego", Poddziałanie nr 2.3.1 „Cyfrowe udostępnienie informacji sektora publicznego ze źródeł administracyjnych i zasobów nauki".

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Rozwoju Regionalnego.

Jest to projekt partnerski, realizowany w siedmiu uniwersytetach medycznych:
Uniwersytet Medyczny im. Piastów Śląskich we Wrocławiu – lider projektu,
Uniwersytet Medyczny w Białymstoku,
Gdański Uniwersytet Medyczny,
Śląski Uniwersytet Medyczny w Katowicach,
Uniwersytet Medyczny w Lublinie,
Pomorski Uniwersytet Medyczny w Szczecinie,
Warszawski Uniwersytet Medyczny
oraz w Instytucie Medycyny Pracy im. prof. Jerzego Nofera w Łodzi.

Fundusze Europejskie Polska Cyfrowa

Unia Europejska Europejski Fundusz Rozwoju Regionalnego

Dane badawcze

Dane badawcze (Research Data)
to zarejestrowane materiały o charakterze faktograficznym, powszechnie uznawane przez społeczność naukową za niezbędne do oceny wyników badań naukowych.

Dane badawcze to zarówno surowe dane, czyli takie, które uzyskano bezpośrednio w wyniku zastosowania narzędzia badawczego oraz takie, które zostały opracowane.

Danymi badawczymi mogą być:

- dane liczbowe,
- dokumenty tekstowe, notatki,
- kwestionariusze, wyniki badań ankietowych,
- nagrania audio i wideo, obrazy,
- zawartość baz danych,
- oprogramowanie,
- wyniki symulacji komputerowych,
- protokoły laboratoryjne,
- opisy metodologiczne.

Otwarte dane badawcze (Open Research Data)
to dane badawcze, które zostały udostępnione w repozytorium lub na innej platformie cyfrowej i do których każdy ma bezpłatny dostęp.

Otwarte dane badawcze można ponownie wykorzystywać, modyfikować i udostępniać z poszanowaniem prawa.

Otwieranie danych badawczych staje się coraz częściej wymogiem stawianym przez instytucje finansujące naukę wobec pracowników naukowych ubiegających się o środki na prowadzenie swoich badań.

- Otwieranie danych pozwala innym naukowcom powtórzyć badania lub je zweryfikować.
- Otwarte dane są gromadzone i udostępniane przeważnie w repozytoriach danych badawczych.
- Nie wszystkie zbiory danych mogą zostać otwarte, w szczególności dotyczy to danych badawczych zawierających dane osobowe, będących podstawą komercjalizacji oraz istotnych dla bezpieczeństwa narodowego.
- Informacja o istnieniu danych zawsze powinna być publicznie dostępna, zapobiegając tym samym ewentualnej duplikacji badań.

Plan zarządzania danymi (Data Management Plan - DMP)

Co to jest plan zarządzania danymi (Data Management Plan – DMP)?

Instytucje i programy finansujące badania naukowe coraz częściej wymagają od naukowców przedstawienia DMP na etapie składania wniosków grantowych (np. Horyzont 2020, NCN).

Plan zarządzania danymi to formalny dokument zawierający zarys tego, co będziemy robić z danymi w trakcie trwania danego projektu badawczego i po jego zakończeniu.

Co powinien zawierać plan zarządzania danymi (DMP)?

- Jakie dane zostaną wytworzone lub zebrane? (rodzaje danych, formaty plików, szacunkowa objętość danych)
- W jaki sposób będą pozyskiwane lub wytwarzane dane? (standardy, metody, oprogramowania, narzędzia)
- Jak zostaną uporządkowane i opisane? (metadane, dokumentacja)
- Kwestie etyczne i prawne (kwestie związane z ochroną danych osobowych, danych niejawnych, etc.)
- W jaki sposób dane zostaną udostępnione? (jak, kiedy, komu)
- Które dane będą przechowywane długoterminowo? (gdzie, jak długo)

Jak przygotować plan zarządzania danymi badawczymi (DMP)?

Można skorzystać z narzędzi do tworzenia planów zarządzania danymi:
DMPTool https://dmptool.org/
DMPonline https://dmponline.dcc.ac.uk/

Przykłady planów zarządzania danymi znajdziemy w wyżej wymienionych narzędziach oraz w serwisie brytyjskiej instytucji specjalizującej się w zarządzaniu danymi badawczymi Digital Curation Centre:
http://www.dcc.ac.uk/resources/data-management-plans/guidance-examples

Wymogi poszczególnych instytucji finansujących naukę mogą być różne.
W przypadku instytucji zagranicznych warto skorzystać z formularzy dostępnych w DMPTool i DMPonline.

Na gruncie polskim należy zapoznać się z opublikowanymi przez Narodowe Centrum Nauki (NCN) *Wytycznymi dla wnioskodawców do uzupełnienia planu zarządzania danymi w projekcie badawczym:*
https://ncn.gov.pl/sites/default/files/pliki/regulaminy/wytyczne_zarzadzanie_danymi.pdf

Przechowywanie i udostępnianie danych badawczych

Przechowywanie danych badawczych (archiwizacja)

Powstające w ramach projektu dane, powinny być przechowywane z zachowaniem zasad bezpieczeństwa, które mają zapobiec ich utracie.

Zastosowane procedury tzn. częstotliwość wykonywania kopii zapasowej, na jakim medium są one zapisywane (pendrive, chmura etc.) oraz gdzie przechowywane są media powinny być opisane w planie zarządzania danymi.

Bezpieczne przechowywanie danych zapewnia zastosowanie reguły 3-2-1:
- zawsze należy mieć ***trzy*** backupy,
- należy używać ***dwóch*** różnych technologii przechowywania danych,
- ***jeden*** backup należy przechowywać w innym miejscu niż dwa pozostałe np. poza uczelnią.

Udostępnianie danych badawczych

W planie zarządzania danymi (DMP) określa się, w jaki sposób dane zostaną udostępnione. Wybierając dane do repozytorium należy kierować się zasadą:
dane powinny być tak otwarte, jak to możliwe i na tyle zamknięte, na ile jest to konieczne.

Dane udostępniane są w postaci pakietów tzw. datasetów czyli zbiorów, które stanowią pewną całość, powiązaną z publikacją, projektem naukowym, eksperymentem.
W datasecie znajdują się wszystkie pliki z danymi oraz pliki niezbędne do ich odczytania bądź interpretacji, np. README.

Zagadnienia do rozważenia przed udostępnieniem danych badawczych

1. Miejsce udostępnienia danych – repozytorium

Polecanymi repozytoriami są:

Polska Platforma Medyczna

http://www.ppm.edu.pl – portal zarządzania wiedzą i potencjałem badawczym, projekt partnerski 7 uniwersytetów medycznych i 1 instytutu badawczego (Uniwersytet Medyczny im. Piastów Śląskich we Wrocławiu – lider projektu, Uniwersytet Medyczny w Białymstoku, Gdański Uniwersytet Medyczny, Śląski Uniwersytet Medyczny w Katowicach, Uniwersytet Medyczny w Lublinie, Pomorski Uniwersytet Medyczny w Szczecinie, Warszawski Uniwersytet Medyczny, Instytut Medycyny Pracy w Łodzi), mający na celu wspieranie współpracy i transferu wiedzy między sektorem nauk medycznych a biznesem. PPM prezentuje osiągnięcia partnerów i potencjał badawczy.

Metadane dla danych badawczych w PPM mają własny format dostosowany do potrzeb funkcjonalności całego systemu. Gromadzone dane badawcze nie powinny przekraczać 30 MB. Zaleca się, aby w przypadku większych zbiorów rejestrować tylko metadane wraz z linkiem do właściwego źródła przechowywania danych badawczych.

Zenodo (OpenAIRE – CERN)

https://www.zenodo.org/ – multidyscyplinarne repozytorium, które powstało we współpracy organizacji OpenAIRE i ośrodka CERN, uruchomione w 2013 r. Każdy z zdeponowanych obiektów otrzymuje identyfikator DOI, co ułatwia jego cytowanie.
Zenodo pozwala na publikowanie danych badawczych o maksymalnej objętości 50 GB dla zestawu (dataset). Możliwe jest opublikowanie dowolnej liczby datasetów.

RepOD (Repozytorium Centrum Otwartej Nauki)

https://repod.pon.edu.pl/pl/about – przeznaczony dla tzw. małych danych, powstających w pracach badawczych prowadzonych przez pojedynczych naukowców lub niewielkie zespoły naukowe. RepOD jest prowadzony w Interdyscyplinarnym Centrum Modelowania Matematycznego i Komputerowego Uniwersytetu Warszawskiego (ICM UW).

Więcej repozytoriów można odnaleźć w wyszukiwarkach:

re3data.org – https://www.re3data.org/ – (Registry of Research Data Repositories) globalny rejestr repozytoriów danych badawczych ze wszystkich dyscyplin akademickich. Umożliwia wyszukiwanie i przeglądanie repozytoriów według dziedziny wiedzy, kraju oraz typu danych badawczych.

OpenDOAR – http://v2.sherpa.ac.uk/opendoar/ – międzynarodowa baza indeksująca biblioteki cyfrowe, repozytoria instytucjonalne i repozytoria danych badawczych wysokiej jakości. Umożliwia wyszukiwanie samych repozytoriów jak i przeszukiwanie ich zasobów.

2. Moment udostępnienia danych

Dane należy udostępnić możliwie szybko. NCN (Narodowe Centrum Nauki) wymaga udostępnienia danych najpóźniej w momencie publikacji wyników badań.

3. Formaty plików

Wybierając format zapisu danych badawczych warto ograniczyć się do tzw. formatów otwartych, czyli takich, które można otworzyć przy użyciu darmowego oprogramowania. Dobór formatów plików do archiwizacji:

Tabele: zaleca się użycie plików o formatach: CSV, TSV, SPSS portable
Nie zaleca się plików: XLS, XLSX

Teksty: zaleca się użycie plików w formatach: HTML, RTF, PDF
Nie zaleca się plików: DOC, DOCX

Media: zaleca się użycie plików w formatach: MP4, Flacc
Nie zaleca się plików: QTFF

Obraz: zaleca się użycie plików o formatach: TIFF, JPGE2000, PNG
Nie zaleca się plików: GIF, JPG

Dane uporządkowane: zaleca się użycie plików o formatach: XML, RDF
Nie zaleca się plików: RDBMS

4. Nazewnictwo plików

Nazwy plików powinny mieć charakter opisowy i odzwierciedlać strukturę zbioru danych. Warto, by w nazwie pliku zawarte były informacje o np. nazwie projektu, dacie zebrania lub przetworzenia danych, unikalnym identyfikatorze czy wersji pliku. Odpowiednio nazwany plik/zbiór plików może znacząco ułatwić użytkownikowi korzystanie z danych.

5. Licencje, czyli jak użytkownik może wykorzystać udostępnione dane

Dane badawcze zaleca się udostępniać z zastosowaniem licencji niewyłącznej lub licencji Creative Commons.

Licencja niewyłączna pozwala na upoważnienie więcej niż jednego podmiotu do udostępniania zestawu danych.

Licencje Creative Commons umożliwiają elastyczne określenie zakresu ochrony dzieła przy użyciu zestawienia warunków przedstawionych poniżej.

4 podstawowe warunki licencji CC to:

BY – Uznanie autorstwa. Wolno kopiować, rozprowadzać, przedstawiać i wykonywać objęty prawem autorskim utwór oraz opracowanie na jego podstawie utwory zależne pod warunkiem, że zostanie przywołane nazwisko autora pierwowzoru.

SA – Na tych samych warunkach. Wolno rozprowadzać utwory zależne jedynie na licencji identycznej do tej, na jakiej udostępniono utwór oryginalny.

NC – Użycie niekomercyjne. Wolno kopiować, rozprowadzać, przedstawiać i wykonywać objęty prawem autorskim utwór oraz opracowanie na jego podstawie utwory zależne jedynie do celów niekomercyjnych.

ND – Bez utworów zależnych. Wolno kopiować, rozprowadzać, przedstawiać i wykonywać utwór jedynie w jego oryginalnej postaci – tworzenie utworów zależnych nie jest dozwolone.

Zestawienia (wariantów) licencji Creative Commons (CC)

CC0 (przekazanie do Domeny Publicznej)
możesz: zwielokratniać, zmieniać, rozpowszechniać i wykonywać utwór, nawet w celu komercyjnym bez pytania o zgodę;

CC BY (uznanie autorstwa)
możesz: kopiować, rozprowadzać, zmieniać, przedstawiać i wykonywać, **pod warunkiem** oznaczenia autorstwa;

CC BY-SA (uznanie autorstwa, na tych samych warunkach)
możesz: kopiować, rozprowadzać, zmieniać, przedstawiać i wykonywać, **pod warunkami:** oznaczenia autora oryginału oraz udzielenia na utwór zależny takiej samej licencji; przykład: Wikipedia i projekty siostrzane;

CC BY-NC (uznanie autorstwa, użycie niekomercyjne)
możesz: kopiować, rozprowadzać, zmieniać, przedstawiać i wykonywać, **pod warunkami**: oznaczenia autora oryginału oraz wykorzystywania go do celów niekomercyjnych (nie wolno na nim zarabiać); utwory zależne mogą być objęte inną licencją;

CC BY-ND (uznanie autorstwa, bez utworów zależnych)
możesz: kopiować, rozprowadzać, przedstawiać i wykonywać, **pod warunkami**: oznaczenia autora oryginału oraz zachowania go w oryginalnej postaci (bez utworów zależnych);

CC BY-NC-SA (uznanie autorstwa, użycie niekomercyjne, na tych samych warunkach)
możesz: kopiować, rozprowadzać, zmieniać, przedstawiać i wykonywać, **pod warunkami**: oznaczenia autora oryginału, nie zarabiania na nim i udostępniania utworów zależnych na tej samej licencji;

CC BY-NC-ND (uznanie autorstwa, użycie niekomercyjne, bez utworów zależnych)
możesz: kopiować, rozprowadzać, zmieniać, przedstawiać i wykonywać, **pod warunkami**: oznaczenia autora oryginału, nie zarabiania na nim i zachowania go w oryginalnej postaci;

Więcej na stronie Creative Commons Polska:
https://creativecommons.pl/poznaj-licencje-creative-commons/

6. Metadane, czyli dane o danych

Metadane stanowią zestaw informacji o konkretnym zbiorze danych badawczych. Mogą zawierać m.in.:

- dane o autorze/autorach,
- streszczenie,
- słowa kluczowe,
- datę publikacji,
- datę powstania zbioru,
- język danych,
- nazwę projektu, itp.

W większości przypadków repozytoria korzystają z własnych standardów metadanych. Podczas deponowania zbioru danych konieczne będzie uzupełnienie formularzy o konkretne metadane.

7. Zasady FAIR Data

Zasady FAIR to wytyczne określające kryteria, które powinny spełniać udostępniane dane, by możliwe było ich ponowne wykorzystanie przez ludzi oraz maszyny. Zasady te dotyczą także metadanych.

FAIR jest akronimem od: Findable – łatwe do znalezienia i wyszukania. Accessible – dostępne dla wszystkich. Interoperable – interoperacyjne, możliwe do zintegrowania np. z innymi zestawami danych. Reusable – wielokrotnego użytku.

Findable – łatwe do znalezienia i wyszukania.

- Czy dane opatrzone zostaną metadanymi?
- Czy będą opisane zgodnie z przyjętymi standardami?
- Czy dane będą posiadać trwałe identyfikatory (DOI)?
- Czy (meta)dane będą zamieszczone lub indeksowane w serwisie, którego zasoby można przeszukiwać?

Accessible – dostępne dla wszystkich.

- Które dane zostaną udostępnione w sposób otwarty?
- Jeśli część danych nie może zostać udostępniona, to dlaczego?
- Czy w takiej sytuacji udostępnione zostaną metadane?
- W jaki sposób i gdzie dane zostaną udostępnione?
- Czy warunki dostępu będą jasno określone?
- Jakie oprogramowanie w dostępie do danych?
- Dane, które ze względu na ochronę prywatności nie mogą zostać opublikowane całkowicie, mogą spełniać wszystkie zasady FAIR.

Interoperable – interoperacyjne, możliwe do zintegrowania np. z innymi zestawami danych.

- Czy przetwarzanie danych będzie możliwe za pomocą otwartego oprogramowania?
- Jaki będzie format plików?
- Czy będzie możliwa wymiana i ponowne wykorzystanie danych przez inne osoby pochodzące z innych instytucji oraz państw?
- Czy możliwe będzie połączenie danych z innymi zbiorami pochodzącymi z innych źródeł?

Reusable – wielokrotnego użytku.

- Czy dane zostaną opatrzone licencją, która pozwoli na ich ponowne wykorzystanie w stopniu tak szerokim jak to możliwe (najszersza jest licencja CC BY)?
- Kiedy możliwe będzie ponowne wykorzystanie danych?
- Czy dane zostaną objęte karencją (embargo), by umożliwić publikację lub uzyskanie patentu? Jeśli tak, jak długo?

Bibliografia

1. Creative Commons Polska [online]. [Dostęp 19.06.2020].
Dostępny w: https://creativecommons.pl/poznaj-licencje-creative-commons/

2. FAIR Principles [online]. GO FAIR. [Dostęp 19.06.2020].
Dostępny w: https://www.go-fair.org/fair-principles/

3. GRUENPETER N. Warsztaty z zarządzania danymi badawczymi. Łódź, 11.06.2019 [online]. [Dostęp 19.06.2020].
Dostępny w: http://otwarty.umed.pl/wp-content/uploads/2019/06/otwarte-dane-badawcze-DMP.pdf

4. GRYGOROWICZ A., MILEWSKA A., WIŚNIEWSKA N. Otwarte dane medyczne [online]. I Kongres Bibliotek Szkół Wyższych, 2019. [Dostęp 19.06.2020]. Dostępny w: https://1kbsw.p.lodz.pl/wp-content/uploads/2019/07/2019-07-Grygorowicz-M-W.pdf

5. Guidelines on FAIR Data Management in Horizon 2020 [online]. European Commission, Directorate-General for Research & Innovation, 2016. [Dostęp 19.06.2020]. Dostępny w: https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

6. Guiding you in Open Science [online]. OpenAIRE [Dostęp 06.07.2020].
Dostępny w: https://www.openaire.eu/guides

7. HOFFMAN-SOMMER M., STARCZEWSKI M., SIEWICZ K. Otwarty dostęp w sektorze publicznym [online]. [Dostęp 19.06.2020]. Dostępny w: https://www.polskacyfrowa.gov.pl/media/27973/WarsztatyPOPC_Otwartydostewsektorzepublicznym_24102016.pdf

8. NOWOCIEŃ T., ROGOWSKA E. Data management plan (DMP) w bibliotece naukowej. Nowe zadania i narzędzia. Medical Library Forum [online]. 2018;11(1):25-30. [Dostęp 19.06.2020]. Dostępny w: https://doi.org/10.34738/mlf.0009

9. Otwarta Nauka [online]. Platforma Otwartej Nauki [Dostęp 06.07.2020].
Dostępny w: https://otwartanauka.pl/

10. ROŻNIAKOWSKA-KŁOSIŃSKA, M. Otwarte dane badawcze w warsztacie pracy naukowca. Biuletyn EBIB [online]. 2018,nr 6 (183). [Dostęp 19.06.2020].
Dostępny w: http://ebibojs.pl/index.php/ebib/article/view/38

11. SZUFLITA-ŻURAWSKA M. Plan Zarządzania Danymi [online]. [Dostęp 19.06.2020].
Dostępny w: https://cdn.mostwiedzy.pl/c0/a3/7d/50/0_20200214103630150545 1_FME/plan-zarzadzania-danymi-2020.pdf