Zarządzanie danymi badawczymi

Poradnik dla naukowców

Oddział Zarządzania Danymi Badawczymi i Otwartej Nauki
BIBLIOTEKA UNIWERSYTETU GDAŃSKIEGO

Cel poradnika

Celem niniejszego poradnika jest wyjaśnienie kluczowych aspektów zarządzania danymi badawczymi i ich praktycznego zastosowania w pracy badawczej. Podręcznik zawiera szczegółowe instrukcje i rekomendacje ułatwiające efektywne zarzadzanie danymi i wdrażanie najlepszych praktyk w tym obszarze. W poradniku można znaleźć m.in. wskazówki dotyczące tworzenia Planu Zarządzania Danymi Badawczymi wymaganego w konkursach NCN, informacje ułatwiające udostępnienie danych badawczych w repozytoriach a także praktyczne narzędzia i rozwiązania usprawniające pracę naukową w obrębie danych.

Słownik pojęć

Dane badawcze - informacje sektora publicznego utrwalone w postaci elektronicznej, inne niż publikacje naukowe, które zostały wytworzone lub zgromadzone w ramach działalności naukowej w rozumieniu art. 4 ustawy z dnia 20 lipca 2018 r. – Prawo o szkolnictwie wyższym i nauce (Dz. U. z 2021 r. poz. 478, 619 i 1630) i są wykorzystywane jako dowody w procesie badawczym lub służą do weryfikacji poprawności ustaleń i wyników badań.[1]

Dane badawcze to wszelkiego rodzaju informacje (ilościowe i jakościowe) wytwarzane, zbierane, przetwarzane, poddawane analizie w ramach projektu badawczego, obejmujące wszystkie możliwe formy zarówno cyfrowe jak i nie cyfrowe.

Data steward - to osoba odpowiedzialna za zarządzanie danymi badawczymi, w tym za ich organizację, dokumentację, zapewnienie jakości oraz przestrzeganie zasad FAIR (Findable, Accessible, Interoperable, Reusable). Data steward wspiera badaczy w procesie planowania zarządzania danymi, ich przechowywania, udostępniania oraz archiwizacji, dbając, aby były zgodne z politykami otwartego dostępu i wymaganiami instytucji finansujących badania.

Repozytorium danych badawczych – narzędzie informatyczne służące do deponowania, przechowywania i udostępniania danych badawczych w taki sposób, aby każdy mógł mieć do nich dostęp w miejscu i czasie przez siebie wybranym na warunkach otwartej licencji Creative Commons.

[1] Ustawa z dnia 11 sierpnia 2021 r. o otwartych danych i ponownym wykorzystywaniu informacji sektora publicznego

Metadane - ustrukturyzowane informacje służące do opisu publikacji naukowych i danych badawczych, pozwalające na ich jednoznaczną identyfikację i umożliwiające ich wyszukiwanie.

Plan zarządzania danymi - (z ang. Data Management Plan - DMP) dokument opisujący zasady zarządzania danymi podczas i po zakończeniu badań. Plan Zarządzania Danymi opisuje rodzaje danych pozyskiwanych, wykorzystywanych lub wytwarzanych w trakcie badań oraz zasady ich udostępniania i długoterminowej archiwizacji.

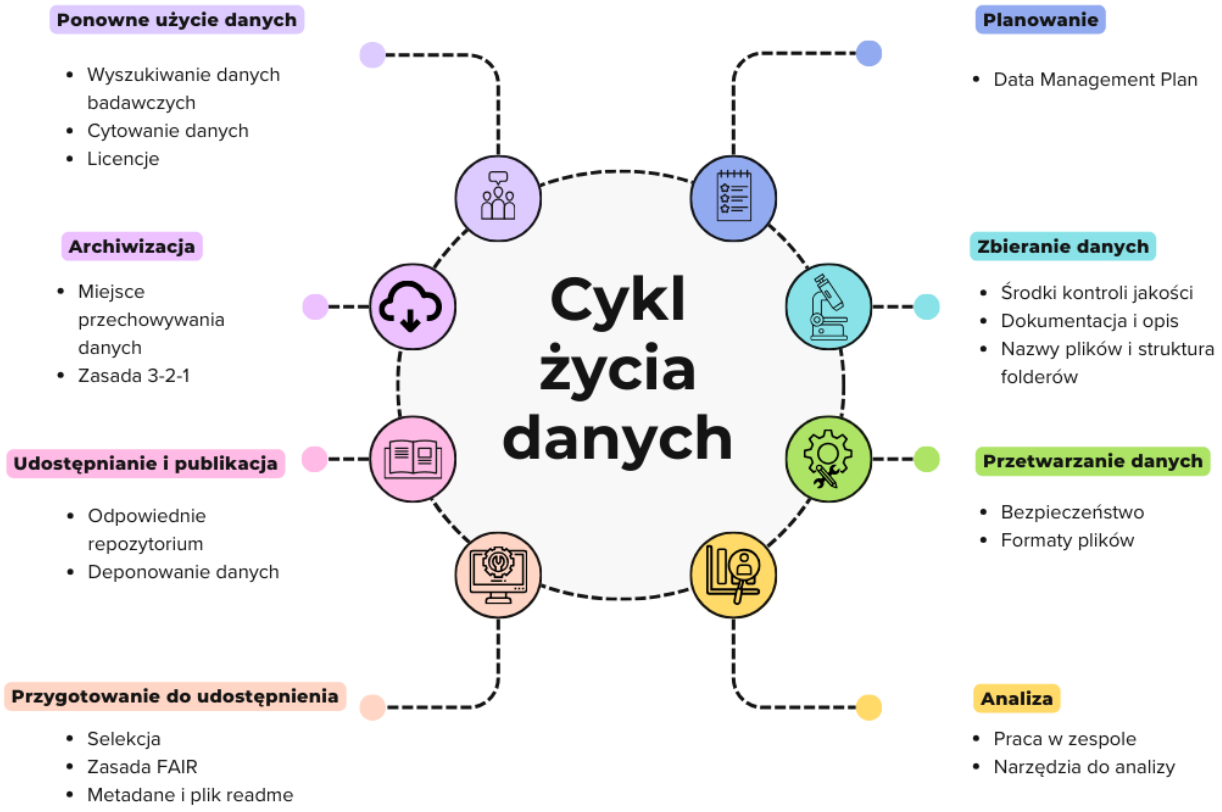
Zarządzanie danymi badawczymi

W tym rozdziale:

- Cele zarządzania danymi badawczymi,
- Środki kontroli jakości w procesie badawczym
- Platforma OSF wspomagająca proces zarządzania danymi (instrukcja)

Dlaczego zarządzanie danymi przez naukowców realizujących badania naukowe jest ważne?

Odpowiednie zarządzanie danymi badawczymi pozwala oszczędzić się czas i pieniądze, uniknąć błędów, utraty danych i bałaganu.[2] Dzięki zarządzaniu proces naukowy jest lepiej i efektywniej zorganizowany co pozwala na optymalizację pracy oraz stanowi gwarancję jakości. Zarządzanie danymi badawczymi obejmuje cały cykl życia danych – od planowania badań po archiwizację i ponowne użycie danych.



[2] I. Bouchrika, What is Research Data Management (RDM) in 2024?, https://research.com/education/research-data-management#benefits [dostęp: 11.06.2024]

Kontrola jakości jako element zarządzania danymi badawczymi

Jednym z elementów procesu zarządzania danymi badawczymi jest kontrola jakości prowadzonych badań. Kontrola jakości danych przekłada się na ostateczne wyniki badań, pozwala na optymalizację i wzrost produktywności pracy. Jest związana z konkretnymi procedurami, stosowanymi narzędziami i procesami badawczymi.[3] NCN w Planie Zarządzania Danymi Badawczymi wymaga m.in przedstawiania środków kontroli jakości stosowanych w projekcie.

Przykłady środków kontroli jakości (narzędzi i procedur) dla nauk ścisłych i eksperymentalnych:

- ✓ EQIPD Core Requirements: https://eqipd-toolbox.paasp.net/wiki/Core_Requirements
- ✓ Good Laboratory Practice (GLP) https://www.gov.pl/web/chemical/good-laboratory-practice
- ✓ Standard Operating Procedures (SOP)
- ✓ kalibracja przyrządów w celu sprawdzenia precyzji, błędu systematycznego lub skali pomiaru;
- ✓ dokonywanie wielokrotnych pomiarów, obserwacji lub próbek;
- ✓ wyspecjalizowany personel;
- ✓ kontrola;
- ✓ stosowanie znormalizowanych metod i protokołów do rejestrowania obserwacji, wraz z zapisami formularzy z jasnymi instrukcjami;
- ✓ stosowanie międzynarodowych standardów i certyfikatów
- ✓ procedury np. pobierania próbek
- ✓ narzędzia do kontroli jakości np “ToxRTool” (Toxicological data Reliability Assessment Tool)

Nauki społeczne i humanistyczne:

- ✓ czyszczenie danych – sprawdzanie poprawności, spójności i użyteczności, usuwanie niepożądanych obserwacji, błędów, literówek
- ✓ narzędzia do analizy danych
- ✓ narzędzia do transkrypcji
- ✓ szablony kwestionariuszy, pytania walidacyjne

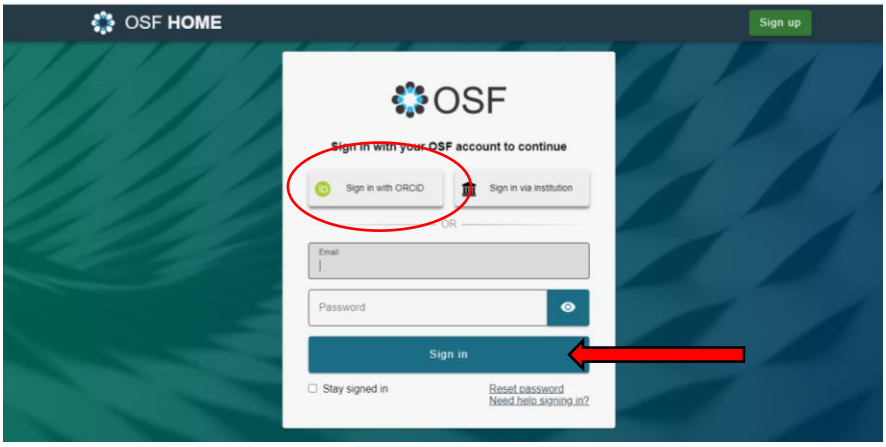
[3] Za: M. Pawłowska, M. Wachowicz, Wprowadzenie do zarządzania danymi naukowymi, Warszawa Difin 2020, s 21.

- ✓ digitalizacja
- ✓ wykazanie wiarygodności źródła i odpowiedni opis zarówno źródła jak i danych

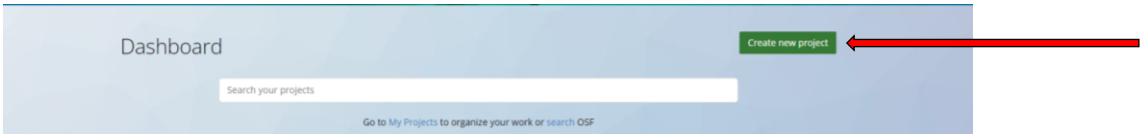
Platforma Open Science Framework

OSF jest bezpłatną i ogólnodostępną platformą wspierającą proces zarządzania danymi badawczymi. Jako narzędzie do współpracy, OSF pomaga zespołom badawczym wspólnie pracować nad projektami, dzielić się wynikami badań wewnątrz zespołu lub udostępniać cały projekt publicznie.

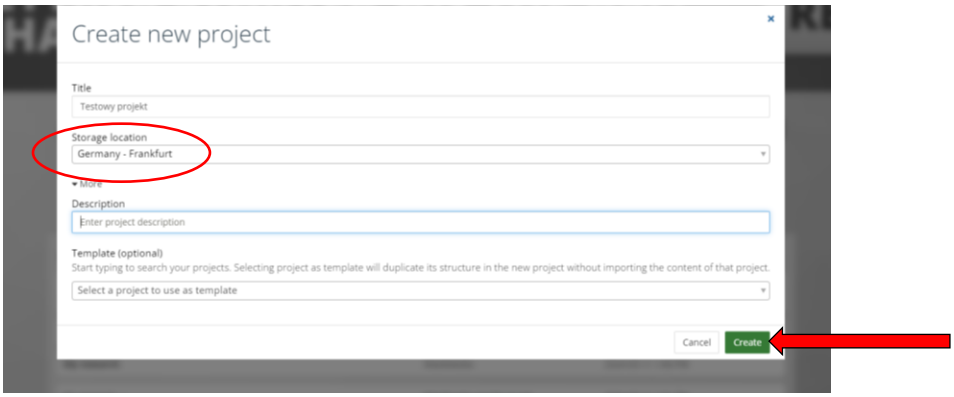
Na platformie można zarejestrować się i zalogować za pomocą konta ORCID.



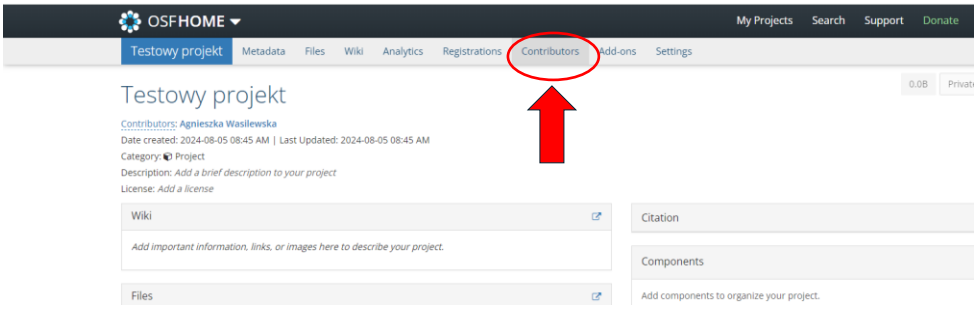
Zaczynamy od kliknięcia zielonego przycisku znajdującego się po prawej stronie ekranu “create new project”, tworzymy nowy projekt.

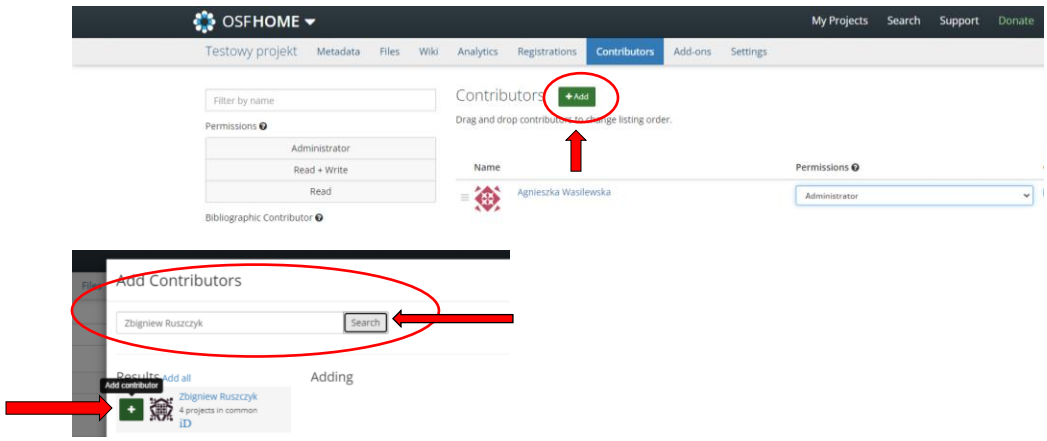


Następnie wybieramy lokalizację. Do wyboru są cztery regiony związane z różnymi przepisami dotyczącymi przechowywania danych badawczych (należy wybrać Niemcy ze względu na to że obowiązują nas te same przepisy UE).

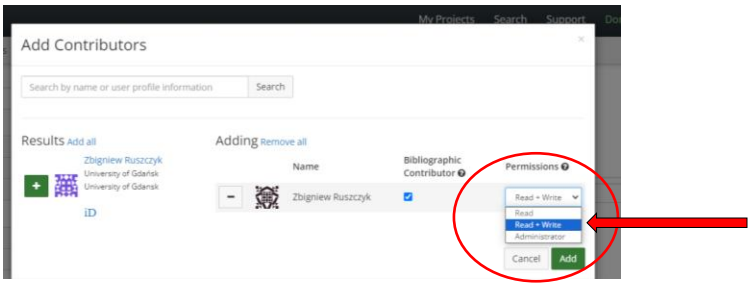


Aby dodać współpracowników należy wybrać przycisk “contributors”. Osoby te muszą wcześniej zarejestrować się na platformie.

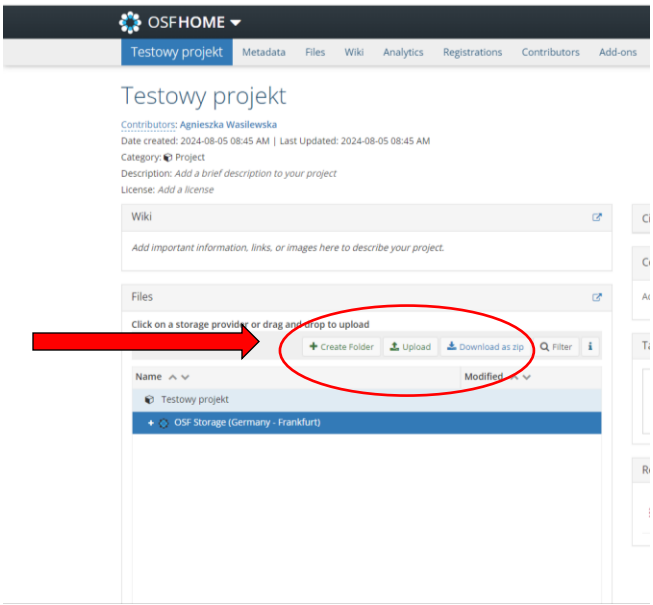


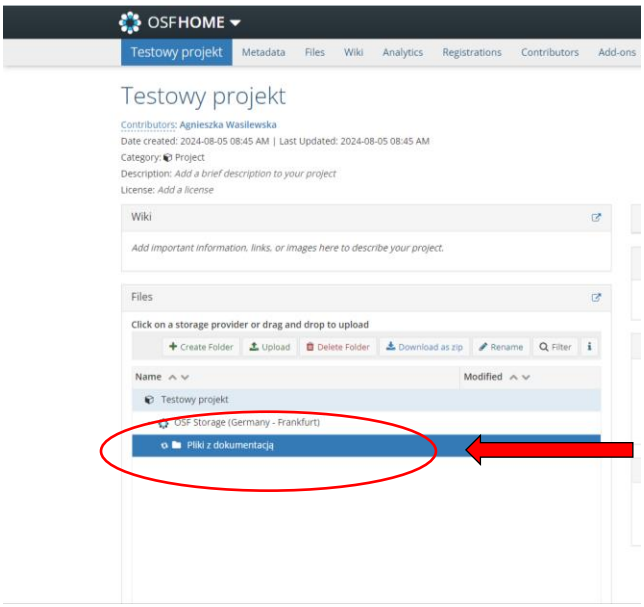


Po wybraniu współpracownika można określić jego uprawnienia w projekcie (tylko odczyt, czy odczyt i edycja lub administrator).

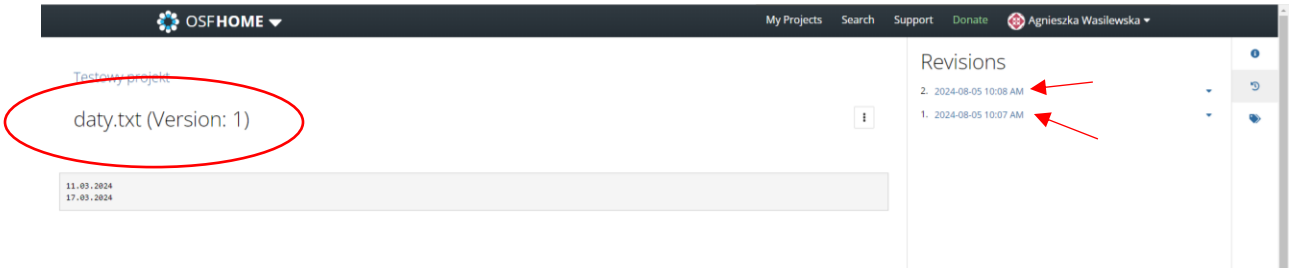


W projekcie możemy zakładać foldery i dodawać pliki. Można załadować do 5 GB dla projektów prywatnych i 50 GB dla publicznych.

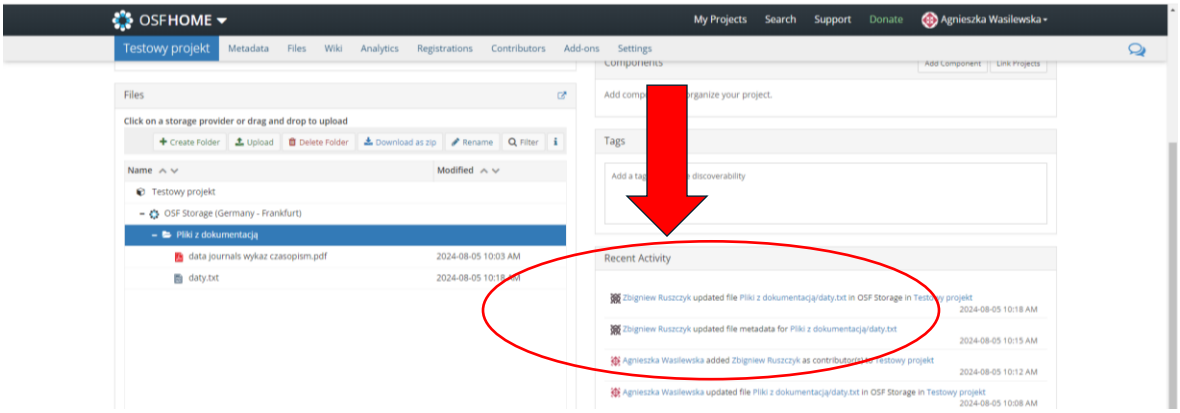




OSF wspiera wersjonowanie plików. Jeśli wgramy plik o tej samej nazwie w którym dodaliśmy zmiany stworzy się automatycznie druga wersja tego pliku widoczna po wejściu w plik.

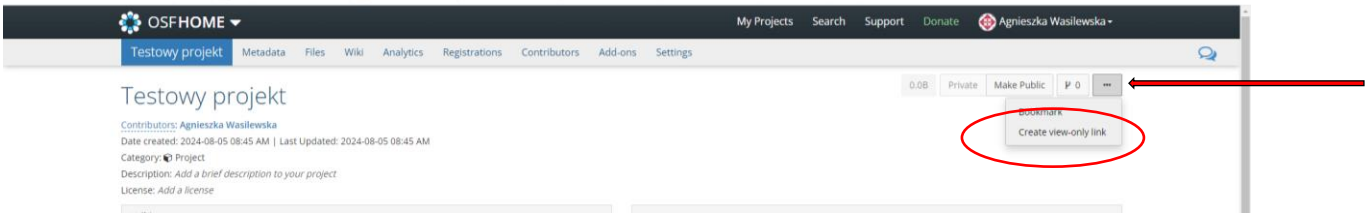


Aktywność poszczególnych członków projektu można śledzić w panelu “Recent activity” po prawej stronie ekranu:

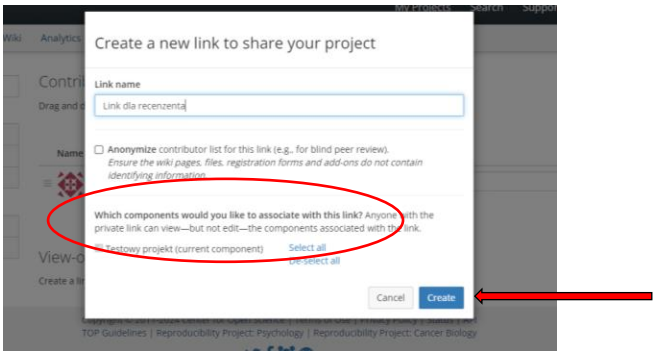


Administrator otrzymuje również powiadomienie mailowe o zmianach w plikach.

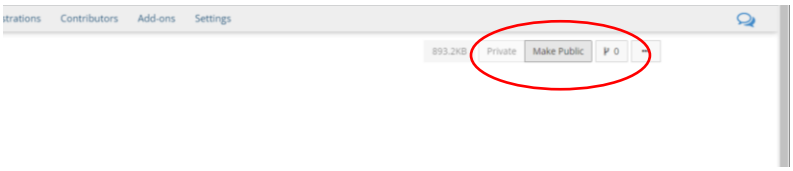
Nasz projekt jest domyślnie ustawiony jako prywatny. W każdej chwili możemy zmienić jego widoczność na publiczną lub wygenerować prywatny link dla konkretnej osoby.

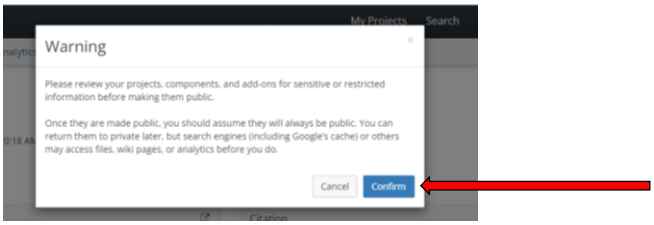


Projekt pozostaje wówczas niewidoczny a jedynie osoba która otrzyma link ma do niego wgląd. Można w tym miejscu zadecydować które element projektu chcemy pokazać.

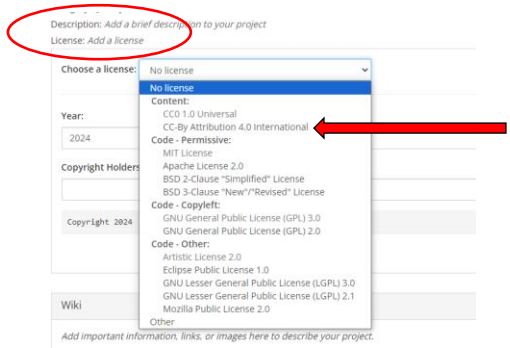


OSF może spełniać również funkcje repozytorium. Możemy uczynić nasz projekt widocznym publicznie i tym samym udostępnić wyniki naszej pracy na zasadach otwartego dostępu.

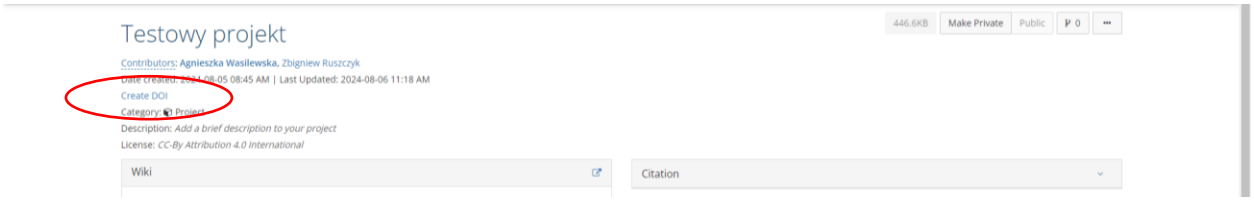


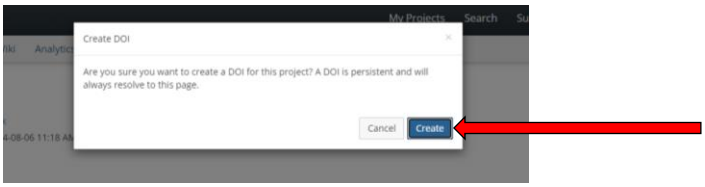


Można również wybrać licencję na jakiej chcemy udostępnić zbiór.

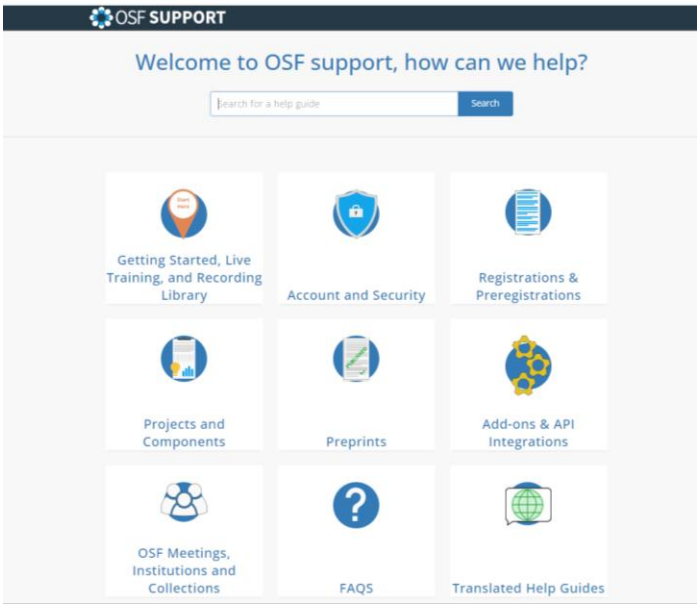


Publicznemu projektowi można przypisać trwały identyfikator DOI.

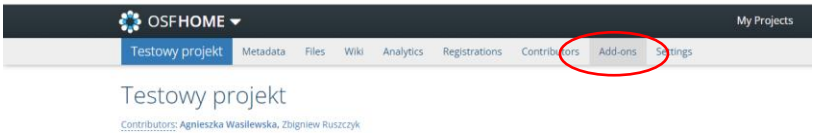


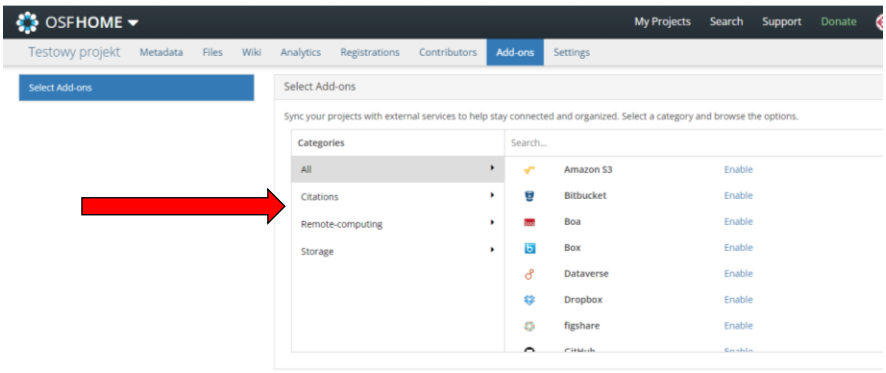


W sekcji “Support” dostępnych jest wiele instrukcji i filmów wyjaśniających możliwości platformy.



OSF można połączyć z dostępnymi narzędziami i kompatybilnym oprogramowaniem np. do przechowywania czy zarządzaniem bibliografią. Połączenie z chmurą OneDrive lub Google Drive daje dodatkową przestrzeń przechowywania, ułatwia również proces dodawania plików.




OSFHOME
My Projects
Search
Support
Donate
Testowy projekt
Metadata
Files
Wiki
Analytics
Registrations
Contributors
Add-ons
Settings
Select Add-ons
Select Add-ons
Sync your projects with external services to help stay connected and organized. Select a category and browse the options.
Categories
Search...
All
Citations
Remote-computing
Storage
Amazon S3
Bitbucket
Boa
Box
Dataverse
Dropbox
figshare
Enable

Wybrane narzędzia wspomagające zarządzanie danymi

Lista bezpłatnych narzędzi, specyficznych dla różnych dyscyplin, wspomagających zarządzanie danymi badawczymi.

Nazwa narzędzia	Opis narzędzia	Link do pobrania
Mendeley Reference Manager for Desktop	Oprogramowanie instalowane na komputerze do zarządzania bibliografią i organizowania cytowań w pracy naukowej, w tym do automatycznego generowania cytatów i bibliografii	https://www.mendeley.com/download-reference-manager/windows
OpenRefine	Bezpłatne, opensource oprogramowanie instalowane na komputerze do oczyszczania, transformacji i formatowania danych tabelarycznych	https://openrefine.org/
IrfanView	Oprogramowanie bezpłatne dla instytucji edukacyjnych służące do przeglądania i konwersji plików graficznych pomiędzy różnymi formatami, w tym: JPEG, JP2000, PNG, TIF, WebP, BMP, PDF, RAW, EMF, itp.	https://www.irfanview.com/
eLabFTW	Darmowy i otwartoźródłowy elektroniczny dziennik laboratoryjny, instalowany na serwerach w technologii Docker	https://www.elabftw.net/
Amnesia	Darmowe narzędzie OpenAire do anonimizacji danych badawczych przy minimalnej utracie jakości informacji	https://amnesia.openaire.eu/download.html
VeraCrypt	Darmowe, opensource oprogramowanie szyfrujące, umożliwiające m. in. szyfrowanie folderów, całych partycji oraz dysków twardych	https://veracrypt.fr/
DirHash	Oprogramowanie na licencji BSD instalowane na komputerze do obliczania sum kontrolnych dla katalogów i plików.	https://idrassi.github.io/DirHash/

	Dostępne algorytmy to m.in. MD5, SHA1, SHA256, SHA512	
GSpot	Oprogramowanie instalowane na komputerze do wykrywania, jaki kodek jest niezbędny do odtworzenia danego pliku audio/video	https://www.headbands.com/gspot/

Planowanie w procesie zarządzania danymi

W tym rozdziale:

- Cele planowania
- Podpowiedzi do poszczególnych punktów DMP (Data management Plan) zgodne z wymogami, które należy spełnić składając wnioski grantowe w konkursach NCN

Cele planowania w procesie zarządzania danymi:

DMP wymagają agencje finansujące badania tj NCN czy Komisja Europejska (Horyzont).

Plan powinien zawierać wszystkie informacje o tym co będzie się działo z danymi badawczymi podczas badań. Plan nie powinien być skopiowany z innego wniosku, czy przygotowany z ogólnikowych odpowiedzi. Na jego podstawie powinno się zrozumieć na czym polegają badania a także w jaki sposób zapewni się ich jakość i wiarygodność.

Plan pozwala zorganizować pracę w czasie, podzielić odpowiednio role i uniknąć poważnych błędów. Dzięki rozpoznaniu potencjalnych zagrożeń można ustalić zabezpieczenia i procedury w celu złagodzenia tych zagrożeń.

Plan może służyć również jako narzędzie do przekazywania informacji o stosowanych procedurach, zasadach i strategiach dla członków zespołu badawczego czy partnerów projektu.[4] Dzięki temu proces badawczy staje się przejrzysty a współpraca bardziej efektywna. Plan zarządzania danymi pomaga utrzymać dokładność, spójność i wiarygodność danych projektu.[5]

Plan zarządzania danymi badawczymi powinien być żywym dokumentem, aktualizowanym w miarę postępowania projektu.

W przygotowaniu DMP wymaganego przy składaniu wniosku pomogą poniższe podpowiedzi.

[4] E. Wilson, The Crucial Role of a Data Management Plan in Project Success, https://www.effectiveprojectmanager.org/blog/the-crucial-role-of-a-data-management-plan-in-project-success, [dostęp 17.06.2023]

[5] CESSDA Training Team (2017 - 2022). *CESSDA Data Management Expert Guide*. Bergen, Norway: CESSDA ERIC. Retrieved from https://dmeg.cessda.eu/ [dostęp: 12.07.2023]

Podpowiedzi do poszczególnych punktów DMP[6]:

1.1 Sposób pozyskiwania i opracowywania nowych danych i/lub ponownego wykorzystania dostępnych danych

Opcja 1. Nowe dane

Jak te dane będą pozyskiwane, jak będą pobierane, opis metodologii, użytego sprzętu, zastosowanych procesów.

- Opisz procesy pozyskiwania danych np.: dane były zbierane podczas eksperymentów laboratoryjnych
- Wykorzystana metodologia np.: metody ilościowe, jakościowe
- Za pomocą jakiej aparatury/narzędzi będą zbierane np.: Mikroskopy i spektrometry

Opcja 2. Ponowne wykorzystanie istniejących danych

Należy wykazać, że dane pochodzą z wiarygodnych źródeł. Czy mamy prawa do ponownego użycia danych?

- Jakie dane będą zbierane
- Źródło danych
- Czy można je ponownie wykorzystać
- Właściciel praw i warunki ponownego wykorzystania
- Zgody na ponowne wykorzystanie danych

1.2 Pozyskiwane lub opracowane dane (np. rodzaj, format, ilość)

Rodzaj:

- dane liczbowe,
- dokumenty tekstowe,
- notatki,
- kwestionariusze,
- wyniki ankiet,

[6] Na podstawie „Wytycznych dla wnioskodawców do uzupełnienia planu zarządzania danymi w projekcie badawczym” https://www.ncn.gov.pl/sites/default/files/pliki/regulaminy/wytyczne_zarzadzanie_danymi.pdf [dostęp: 24.11.2022]

- transkrypcje
- nagrania audio
- video,
- fotografie,
- zawartość baz danych,
- oprogramowanie,
- wyniki symulacji komputerowych,
- protokoły metodologiczne,
- obserwacje laboratoryjne
- Inne:

Objętość:

- Szacunkowa wielkość danych .....MB / GB / TB
- Wskaż proporcje danych surowych (nieprzetworzonych), danych przetworzonych i innych wyników wtórnych (np. raportów).

Format:

W jakim formacie będą znajdować się dane - należy stosować formaty otwarte i standardowe.

- Dane będą w formatach otwartych
- Dane nie będą zapisane w formatach otwartych (Uzasadnij dlaczego):

Zalecane formaty danych:

Tabele:	.csv, SPSS (.dat/.sps), Stata, .tab
Dane tekstowe:	.rtf, .txt, .xml., .odt
Media:	.aiff, .flac, .mp4,
Obrazy:	TIFF, JPEG2000, PNG.

Dokumentacja:	.rtf, PDF/A, .pdf, HTML, .odt

2.1 Metadane i dokumenty (np. metodologia lub pozyskiwanie danych oraz sposób porządkowania danych) towarzyszące danym

Standardy metadanych – dzięki temu dane będą możliwe do odnalezienia nie tylko przez innych naukowcó ale i przez algorytmy wyszukiwawcze. Repozytoria danych badawczych posiadają odpowiednie formatki do tworzenia metadanych. Dane umieszczone w uznanym, dziedzinowym repozytorium danych badawczych będą posiadały opis metadanych zgodny ze standardem.

- Standard narzucony przez repozytorium
- Dublin Core
- DataCite Metadata Schema
- Standard dziedzinowy (jaki):
- Darwin Core
- DDI
- EBU
- EAD
- METS
- INSPIRE
- MODS
- TEI
- VRA
- CIF

Dokumentacja - opisz typy dokumentacji, które będą towarzyszyć danym, aby pomóc użytkownikom w ich zrozumieniu i ponownym użyciu. Dokumentacja powinna opisywać ogólny kontekst zebranych danych

Co będzie zawierała dokumentacja opisująca dane:

- informacje dotyczące źródła pochodzenia danych,
- informacje o metodologii zbierania danych,
- zastosowanych procedurach przy zbieraniu danych,

- definicje zmiennych czy jednostek pomiarowych.

Czy będzie dołączony **plik readme** opisujący dane?

Organizacja folderów i plików – stosowanie spójnej konsekwentnej struktury, nazewnictwo plików odzwierciedlające zawartość, standardowy format daty – tak aby łatwo można było odnaleźć i posortować zawartość.

- W jaki sposób będą opisywane pliki i foldery
- Czy będą używane słowniki kontrolowane do opisu danych

2.2 Stosowane środki kontroli jakości danych

1. Dane wytworzone w trakcie badań naukowych:

Stosowane procedury – rodzaj środków kontroli wytwarzania/zbierania danych i ich częstotliwość. Należy opisać, w jaki sposób planuje się kontrolować i dokumentować jakość i spójność danych np.

- kalibracja sprzętu,
- wielokrotne pomiary,
- znormalizowane metody wraz z formularzami,
- szablony kwestionariuszy,
- pytania kontrolne
- czyszczenie danych
- programy do analizy danych/ do zarządzania danymi
- SOP (Standard Operating Procedures)
- GLP (dobra praktyka laboratoryjna)
- ISO
- audyt
- wyszkolony personel
- procedury korzystania ze sprzętu
- inne:

2. Ponownie wykorzystane dane pochodzące z innych źródeł:

Należy wykazać, że dane będą pochodziły z wiarygodnego źródła i będą odpowiednio udokumentowane.

3.1 Przechowywanie i tworzenie kopii zapasowych danych i metadanych podczas badań

Gdzie będą przechowywane dane i kopie zapasowe podczas badań?

- Chmura chroniona umową dwustronną z uczelnią - np. OneDrive (Microsoft 360 dla UG)
- Serwer uczelniany
- Dedykowana macierz
- Dysk zewnętrzny (Uwaga: istnieje ryzyko uszkodzenia, zabezpiecz dane dodatkowo w chmurze)
- Usługa TASK dla projektów NCN (przechowywanie 100 GB + 200 GB na back-up)

! Nie zaleca się pen drive (może zostać uszkodzony, zgubiony, ukradziony)

! Nie zaleca się używania prywatnych komputerów i laptopów

! Nie zaleca się używania chmur komercyjnych (np. Google Drive, Dropbox itp.)

Kopie zapasowe

Stosuj zasadę 3-2-1 (co najmniej 3 kopie, na dwóch różnych nośnikach w tym co najmniej jeden w innej lokalizacji)

- Należy opisać jak często będzie wykonywana kopia zapasowa
- Czy będzie to proces automatyczny/ręcznie kontrolowany, gdzie i przez kogo?

3.2 Sposób zapewniania bezpieczeństwa danych oraz ochrony danych wrażliwych podczas badań

Kto będzie miał dostęp do danych podczas trwania projektu, jak dostęp będzie kontrolowany?

Czy dane zawierają dane wrażliwe? Np.:

- dane ujawniające pochodzenie rasowe lub etniczne,
- dane ujawniające poglądy polityczne,
- dane ujawniające przekonania religijne lub światopoglądowe,
- dane ujawniające przynależność do związków zawodowych,
- dane genetyczne,
- dane biometryczne (wykorzystywane w celu jednoznacznego zidentyfikowania osoby fizycznej),
- dane dotyczące zdrowia,
- dane dotyczące seksualności lub orientacji seksualnej.

Jeśli dane są poufne/wrażliwe należy przedstawić wszelkie odpowiednie środki bezpieczeństwa i odnotować wszelkie formalne standardy, które będziesz przestrzegać np.

- szyfrowanie,
- anonimizacja,
- ściśle kontrolowany dostęp.

Przewidywanie możliwych zagrożeń i ryzyka oraz określenie procedur i sposobów postępowania.

Zabezpieczenie fizyczne – (przed nieuprawnionym dostępem, zniszczeniem, uszkodzeniem nośników, kradzieżą).

Zabezpieczenie cyfrowe – (backup, silne hasło, antywirus, firewall).

4.1 Sposób zapewnienia zgodności z przepisami dotyczącymi danych osobowych i bezpieczeństwa danych w przypadku przetwarzania danych osobowych

Czy w badaniu będą przetwarzane dane osobowe (osób innych niż członkowie zespołu)?

NIE – zaznacz nie dotyczy

TAK: Powołujemy się na Politykę ochrony danych w instytucji:

- Inspektor Ochrony Danych UG.

- Polityka ochrony danych osobowych w Uniwersytecie Gdańskim - Dane będą gromadzone i przetwarzane zgodnie z rozporządzeniem o ochronie danych osobowych: Zarządzenie Rektora UG nr 163/R/21 w sprawie Polityki ochrony danych osobowych w Uniwersytecie Gdańskim.
- Polityka bezpieczeństwa teleinformatycznego Uniwersytetu Gdańskiego: zarządzenie nr 37/R/22 Rektora Uniwersytetu Gdańskiego z dnia 24 marca 2022 roku

Od uczestników badania należy uzyskać świadomą zgodę na piśmie na udział w badaniu i przetwarzanie danych.

Dane należy odpowiednio zabezpieczyć (anonimizacja, pseudonimizacja, szyfrowanie)

4.2. Sposób zarządzania innymi kwestiami prawnymi, np. prawa własności Intelektualnej

Wskazanie właścicieli praw do wszelkich pozyskiwanych i wytwarzanych danych.

- Współpraca z inną jednostką lub korzystanie ze sprzętu innej jednostki - należy ustalić z jednostką współpracującą podział praw własności intelektualnej.
- Na jakich zasadach będą udostępnione dane - zaleca się stosowanie standardowych licencji Creative Commons.
- Czy istnieją ograniczenia prawne ponownego wykorzystania istniejących danych - uwzględnienie praw autorskich. Jeśli korzysta się z danych z bibliotek czy archiwów - stosowanie się do zasad i regulaminów.

Aspekt etyczny - czy istnieją przeciwskazania natury etycznej, czy potrzebna jest ocena właściwej komisji np. wydziałowej, uczelnianej komisji etycznej, Komisji Bioetycznej.

Zgody i pozwolenia na prowadzenie badań – czy badania wymagają zezwoleń specjalistycznych organów.

- Właścicielem danych wytworzonych będzie UG
- Podział praw regulowany umową z jednostką x
- Prawa będą należały do innej jednostki

Dane będą udostępnione na licencji:

- CC-BY

- CC 0 (przekazanie do domeny publicznej)
- Jeśli inna należy uzasadnić dlaczego:

5.1 Sposób i termin udostępniania danych. Ewentualne ograniczenia w udostępnianiu danych lub przyczyny embarga

Przykładowe repozytoria danych badawczych:

Repozytoria ogólnego przeznaczenia	**Instytucja prowadząca**	**Informacje dotyczące plików**	**Istotne informacje**
Repod	ogólnopolskie repozytorium prowadzone przez ICM. UG posiada kolekcję instytucjonalną	plik do 5 GB, bez limitu ilość plików	bezpłatne, nadaje DOI, metadane w schemacie DataCite
MOST	prowadzone przez Politechnikę Gdańską	pliki do 2 GB	posiada certyfikat Core Trust Seal. Nadaje DOI, metadane w DataCite Schema. Należy napisać maila z prośbą o nadanie uprawnień do deponowania: most@pg.edu.pl
Zenodo	europejskie repozytorium stworzone przez OpenAire prowadzone przez CERN	wszystkie typy danych i formaty plików, do 50 GB	metadane pobierane w formacie JSON, schemat Datacite, bezpłatne, DOI

Mendeley data	Elsevier	wszystkie typy danych i formaty plików, do 10 GB	bezpłatne, nadaje DOI
Open Science Framework (OSF)	Centre of Open Science	Wszystkie formaty, do 5 GB dla projektów prywatnych i 50 GB dla publicznych	Bezpłatne, nadaje DOI, można deponować zarówno dane jak i preprinty
Figshare	Digital Science	Pliki do 20 GB, wszystkie formaty	DataCite i Dublin Core, DOI, bezpłatne do 20 GB. Dla większych datasetów płatna usługa figshare+

repozytorium	**dziedzina**	**Rodzaj danych i formaty plików**	**Istotne informacje**
Pangaea	Nauki o ziemi i środowisku	Pliki do 100 MB (większe deponuje się za pomocą specjalnego linku), formaty	nadaje DOI, schemat metadanych Darwin Core i Dublin Core, posiada certyfikat Core Trust Seal, pomoc kuratora danych
Pride	Nauki biologiczne	Bez limitu wielkości plików, akceptowalne formaty Raw, FASTA, baf, fid, dat, MS, SMS, mgf, mzTab, mzIdentML, json, xml, txt, xls, Other, mzML	darmowe, standard metadanych MIBBI, nadaje unikalny

			identyfikator PXD
Repozytorium Danych Społecznych RDS	Nauki społeczne	Dane jakościowe. Formaty: ASCII, CSV,TSV, R, SAV, POR, DTA, SAS, STA, xls, xlsx	Nadaje DOI, darmowe, wymagane dołączenie dokumentacji badania (szablon na stronie)
MX_RDR *Macromolecular Xtallography Raw Data Repository*	Chemia	Preferowany ZIP	Nadaje DOI, DataCite metadata schema,
PDB Worldwide Protein Data Bank	Biologia, chemia	Format PDBx/mmCIF, PDB	Identyfikator PDB ID, standard metadanych mmCIF (Macromolecular Crystallographic Information File)
GBIF Global Biodiversity Information Facility	Nauki przyrodnicze	Nie ma ograniczeń wielkości plików. Można przesyłać dane w formie w plików tekstowych (CSV lub TSV) i plików XML. Publikowanie danych graficznych odbywa się poprzez powiązanie obrazów cyfrowych za pomocą URL z rekordami danych o występowaniu. GBIF nie przechowuje bezpośrednio plików graficznych, plik graficzne muszą być hostowane na zewnętrznych serwerach.	Standardy metadanych: Darwin Core, EML, BioCASe / ABCD Repozytorium posiada cerytfikat WDS. Nadaje identyfikator DOI
HEPData	Nauki fizyczne	Dostępne narzędzia do konwersji plików na stronie	Nadaje DOI

CLARIN-PL	Językoznawstwo	Gromadzi korpusy językowe, materiały tekstowe, wypowiedzi i źródła, zasoby leksykalne, słowniki, listy słów; modele językowe i gramatyki konkretnych języków; narzędzia, aplikacje	Standard metadanych OAI-ORE, Dublin Core, nadaje identyfikator handle
ARCHE	humanistyka	Formaty i metadane: https://arche.acdh.oeaw.ac.at/browser/formats-filenames-and-metadata#formats	Identyfikator handle. Repozytorium posiada certyfikat Core Trust Seal i CLARIN certificate B. Bezpłatne

Kiedy udostępnisz dane? – czy jest embargo (jak długo).

- Dane będą udostępnione możliwie szybko – w momencie publikacji wyników badań.
- Embargo - (należy uzasadnić przyczynę):
- Dane badawcze nie mogą zostać udostępnione (uzasadnienie):

(Ograniczenia w udostępnianiu wynikające z przyczyn prawnych, potencjału komercjalizacji, dane poufne, wrażliwe, tajemnica przedsiębiorstwa itp.)

Jeśli jakieś zbiory nie mogą zostać udostępnione lub dostęp do nich musi zostać ograniczony - należy wyjaśnić, dlaczego jest to konieczne (np. ograniczenia o charakterze prawnym). W takich przypadkach metadane (opisujące dane) powinny być udostępnione w repozytorium.

5. 2 Sposób wyboru danych przeznaczonych do przechowywania oraz miejsce długotrwałego przechowywania (repozytoria, archiwum)

Selekcja:

Jakie dane należy zachować a jakie zniszczyć? Na jakiej podstawie zdecydujesz jakie dane zachować np. jeśli dane mogą zostać ponownie wykorzystane w celu potwierdzenia

wyników badań, przeprowadzenia nowych badań. Czy chcesz udostępniać dane z eksperymentu, który się nie powiódł, żeby inni nie powielali błędów? Zdecyduj, które dane zachować i na jak długo.

Jak będzie wyglądał proces selekcji?

Które dane będą udostępnione?

Nie wszystkie dane muszą być udostępnione w repozytorium – tylko te wyselekcjonowane, wyczyszczone, najlepszej jakości.

Wymóg udostępnienia dotyczy danych stanowiących podstawę publikacji, służące do weryfikacji wyników badań.

Czy będą udostępniane również dane surowe? – jeśli nie to oryginalne surowe dane powinny być zarchiwizowane.

- Udostępnione będą podstawowe dane służące do weryfikacji wyników badań
- Udostępnione będą również dane surowe
- Udostępnione będą wyselekcjonowane dane
- Dane nie będą udostępnione z powodu:

Archiwizacja:

Jak długo będą przechowywane - (w rozumieniu NCN uzasadniony okres przechowywania danych to minimum 10 lat)

5.3 Metody lub narzędzia programowe umożliwiające dostęp do danych i korzystanie z nich

Czy potencjalni użytkownicy danych będą potrzebowali określonych narzędzi aby uzyskać dostęp do danych i ponownie je wykorzystać?

Czy do skanowania lub konwersji niezbędny będzie dodatkowy sprzęt lub oprogramowanie?

Należy opisać oprogramowanie używane w trakcie zbierania danych, narzędzia (np. MatLab).

Zasada FAIR (interoperacyjność) - czy jest możliwe wykorzystanie przez inne osoby i podmioty (w tym maszyny). Jeśli wykorzystanie zbioru wymaga opracowania translatora,

sposobu mapowania czy konwersji formatów, świadczy to o jego ograniczonej interoperacyjności. O pełnej interoperacyjności możemy więc mówić wtedy gdy zdeponowany przez nas zbiór dostępny jest za pośrednictwem API, a jego metadane dodatkowo za pośrednictwem protokołu OAI-PMH

- Do korzystania z danych nie będzie potrzebne dodatkowe oprogramowanie
- W trakcie zbierania danych użyto oprogramowania:
- Oprogramowanie niezbędne do ponownego użycia danych (link):

5.4 W jaki sposób zagwarantują Państwo stosowanie unikalnego i trwale przypisanego identyfikatora (takiego jak cyfrowy identyfikator dokumentu elektronicznego (DOI)) dla każdego zbioru danych?

Identyfikator nadawany jest w repozytorium – przy wyborze repozytorium danych badawczych należy zwrócić uwagę czy dane repozytorium nadaje DOI lub inny unikalny, trwały identyfikator.

Trwały identyfikator - umożliwia dotarcie do zbioru nawet kiedy zmieni się adres url., pozwala na trafną i wydajną lokalizację danych, służy jako odniesienie oraz pozwala na łatwe śledzenie cytowań i wielokrotne użycie.

- Repozytorium nadaje identyfikator DOI
- Repozytorium nadaje inny trwały identyfikator: ARK (Archival Resource Key), PURL (Persistent Uniform Resource Locator), HNDL (Handle), URN (Universal Resource Name), URI (Uniform Resource Identifier), PDB ID
- Inny sposób nadania trwałego identyfikatora:

6.1 Osoba odpowiedzialna za zarządzanie danymi

- Kto będzie odpowiedzialny za każde działanie związane z zarządzaniem danymi i jak długo?
- Jak zostaną rozdzielone obowiązki między strony partnerskie w ramach wspólnych projektów badawczych?

- Przedstaw role i obowiązki dotyczące poszczególnych działań, np. przechwytywanie danych, tworzenie metadanych, jakość danych, przechowywanie i tworzenie kopii zapasowych, archiwizacja i udostępnianie danych. Zastanów się, kto będzie odpowiedzialny za zapewnienie przestrzegania odpowiednich zasad.

Osoba odpowiedzialna za zarządzanie danymi:

- Dedykowany członek zespołu
- Kierownik projektu
- Data steward

6.2 Środki (finansowe i czasowe) przeznaczone do zarządzania danymi

Środki potrzebne do realizacji planu. Mogą one obejmować koszty przechowywania, sprzęt, oprogramowanie, czas pracy personelu, koszty przygotowania danych do depozytu i opłaty repozytorium (przeważnie repozytoria są bezpłatne). Tam, gdzie potrzebne są dedykowane zasoby, należy je nakreślić i uzasadnić konieczność ich użycia.

Jakimi zasobami dysponujemy, a jakie dopiero musimy pozyskać? Jakie będą koszty związane będą z zapewnieniem standardów FAIR w projekcie? (koszty związane z długotrwałym przechowywaniem i udostępnianiem danych).

- Środki na przechowywanie danych
- Korzystamy z bezpłatnych rozwiązań
- Środki na repozytorium (jeśli jest odpłatne)
- Repozytorium bezpłatne
- Zatrudnienie dedykowanego datastewarda

Przechowywanie danych badawczych

W tym rozdziale:

- Stosowanie zasady 3-2-1
- Planowanie zapotrzebowania na przestrzeń dyskową w projekcie
- Bezpieczeństwo danych w trakcie realizacji badań

Stosowanie zasady 3-2-1

Planując a następnie realizując projekt badawczy należy stosować zasadę 3-2-1 w odniesieniu do liczby kopii danych badawczych

- Powinny istnieć **3 kopie** danych, aby zminimalizować ryzyko ich utraty.
- Kopie powinny być przechowywane na **2 różnych nośnikach** (np. lokalnie i w chmurze), aby zmniejszyć ryzyko jednoczesnej awarii obu kopii.
- **1 kopia powinna być przechowywana w innej lokalizacji** fizycznej (np. na serwerach w chmurze lub na serwerze instytucji), by zapobiec utracie danych na skutek katastrof lokalnych.

Planowanie zapotrzebowania na przestrzeń dyskową w projekcie

W trakcie planowania projektu badawczego należy dokonać oszacowania przestrzeni dyskowej oraz miejsc przechowywania wytworzonych w trakcie realizacji projektu danych. W procesie tym należy uwzględnić w szczególności następujące zagadnienia:

- Specyfika formatów danych stosowanych w projekcie w kontekście zajętości przestrzeni dyskowej
- Łączna Ilość i rozmiar plików danych, które zostaną wytworzone w projekcie
- Sposób przechowywania surowych danych w zakresie zastosowania kompresji i oszczędności przestrzeni dyskowej
- Liczba planowanych serii danych, które zostaną wytworzone
- Uwzględnienie przestrzeni dyskowej potrzebnej dla wersjonowania danych, o ile będzie stosowane
- Liczba etapów czyszczenia danych surowych oraz sposób postępowania z danymi zbędnymi, w tym testowymi
- Rozmiar przestrzeni dyskowej dostępny na potrzeby kopii zapasowych
- Liczba i czas wykonywania kopii zapasowych, które będą wykonywane
- Sposób zarządzania retencją kopii zapasowych

Bezpieczeństwo danych w trakcie realizacji badań

W celu zapewnienia bezpieczeństwa danych w trakcie realizacji badań należy stosować się do Polityki bezpieczeństwa teleinformatycznego obowiązującej w UG oraz uwzględnić inne działania wpływające na bezpieczeństwo danych, w tym:

- Ustalenie zasad bezpieczeństwa jakie będą stosowane w trakcie realizacji badań, w tym poziom i sposób zabezpieczenia danych zbieranych z wykorzystaniem urządzeń przenośnych
- Przeszkolenie wszystkich osób uczestniczących w projekcie badawczych w zakresie stosowanych zasad bezpieczeństwa
- Ustalenie struktury katalogów i uprawnień w wykorzystywanych systemach informatycznych adekwatne do ról i odpowiedzialności członków zespołu projektowego
- Ustalenie ról i odpowiedzialności członków zespołu projektowego, w tym
    - Osobę odpowiedzialną za wykonywanie backupów
    - Osobę odpowiedzialną za weryfikację poprawności wykonanych backupów
    - Osobę odpowiedzialną za nadawanie i odbieranie uprawnień
    - Osobę odpowiedzialną za zarządzanie pojemnością przestrzeni dyskowej
- Ustalenie i egzekwowanie na bieżąco zasad prowadzenia dokumentacji projektu, w tym niezbędnych rejestrów, dzienników, raportów oraz opisu danych zgodnie z ustalonym w projekcie standardem

Otwarte dane badawcze

W tym rozdziale:

- Kiedy otwierać dane
- Zasada FAIR
- Kryteria wyboru repozytorium
- Jak odnaleźć odpowiednie repozytorium (instrukcja korzystania z wyszukiwarki re3data)

Dane powiązane z publikacją

Proces deponowania danych badawczych w repozytorium może odbywać się jednocześnie z procesem publikacyjnym. Dane można zdeponować w tzw. wersji roboczej, bez publicznego dostępu. Wówczas dataset nie jest widoczny jednak istnieje możliwość wygenerowania prywatnego linku np. dla recenzentów. Dodatkowo danym zostaje nadany trwały identyfikator (który nie jest jeszcze aktywny) ale można za jego pomocą powiązać zestaw danych z publikacją. Dane badawcze najlepiej jeśli zostają upublicznione zaraz po lub w momencie opublikowania powiązanego artykułu.

Otwarte dane badawcze powinny spełniać wymogi zasady FAIR. Zasada ta określa jakie cechy powinny mieć dane badawcze, aby były możliwe do odnalezienia i przetworzenia zarówno dla ludzi jak i dla algorytmów wyszukiwawczych.

Deklaracje dotyczące dostępności danych (Data availability statements)

Wielu wydawców wymaga dołączenia do publikacji tzw. *Data availability statements*, deklaracji o dostępności danych. W takim oświadczeniu autor określa, czy udostępni dane badawcze które były podstawą publikacji, a jeśli tak, to gdzie i w jaki sposób można uzyskać do nich dostęp. Niektórzy wydawcy podają gotowe formułki deklaracji dostępności, które należy dołączyć do artykułu.[7]

Przykłady deklaracji dostępności:

1. **Publicznie dostępne dane**: Zbiór danych analizowany w tym badaniu jest dostępny w [nazwa repozytorium] pod adresem [DOI/link], na licencji [typ licencji, np. CC BY 4.0].
2. **Dane dostępne na żądanie:** Dane wykorzystane w tym badaniu mogą zostać udostępnione przez autora korespondencyjnego na uzasadnioną prośbę.
3. **Dane o ograniczonym dostępie**: Ze względu na ograniczenia [np. etyczne, prawne, potencjał komercjalizacyjny] dane nie mogą być udostępnione publicznie. Dostęp może zostać przyznany na wniosek i po zatwierdzeniu przez [np. komisję uczelnianą].
4. **Dane objęte embargiem**: Dane powiązane z publikacją są objęte embargiem do dnia [data zakończenia embarga]. Po tym terminie będą dostępne w [nazwa repozytorium] na określonych warunkach licencji.
5. **Ściśle chronione dane**: Ze względu na obowiązujące przepisy prawne, zasady ochrony prywatności lub inne ograniczenia, dane użyte w tym badaniu nie mogą zostać udostępnione publicznie ani na indywidualny wniosek.
6. **Brak nowych danych**: W ramach tego badania nie wygenerowano nowych danych. Wszystkie analizy opierają się na wcześniej opublikowanych źródłach, które zostały cytowane w artykule.

Zasada FAIR

Findable – Odnajdywalne. Metadane i dane powinny być łatwe do odnalezienia zarówno dla ludzi, jak i dla komputerów.

1. **(Meta)danym jest przypisany unikalny i trwały identyfikator** - przykładami identyfikatorów spełniających te warunki są: Digital Object Identifier (DOI) nadawane dla datasetu, Handle System, Archival Resource Key (ARK),

[7] Cambridge University Press, Data availability statements: https://www.cambridge.org/core/services/data-availability-statements [dostęp: 17.02.2025]; Taylor&Francis, Writing a data availability statement: https://authorservices.taylorandfrancis.com/data-sharing/share-your-data/data-availability-statements/ [dostęp: 17.02.2025].

ProteomeExchange Dataset ID (PXD), Uniform Resource Identifier (URI) taki jak URL albo URN (ale pod warunkiem, że jest gwarantowana jego niezmienność w czasie)

2. **Dane są opisane bogatymi metadanymi** - zestaw pól metadanych opisujących dane może być uniwersalny, czyli oparty na ogólnym standardzie metadanych, np. DublinCore (i wtedy należy dataset opisać za pomocą pól takich jak: Tytuł datasetu, identyfikator, Wydawca, data publikacji, Opis lub Podsumowanie, Słowa kluczowe, itp.) lub też może to być standard metadanych specyficzny dla danej dziedziny, np. DDI. Zwykle to repozytorium w którym dane są składowane narzuca minimalny zestaw metadanych te dane opisujących.
3. **Metadane zawierają identyfikator danych, które opisują** - oznacza to, że wśród pól metadanych opisujących dany zbiór danych musi znajdować się unikalny i trwały identyfikator (np. wspomniany DOI , który kieruje do danego zbioru danych.
4. **(Meta)dane są zarejestrowane lub zindeksowane w zasobach, które można przeszukiwać ręcznie jak również z pomocą algorytmów przeszukujących** - aby spełnić ten wymóg należy wybrać do przechowywania zbioru danych repozytorium, które "rozumie" algorytmy przeszukujące zasoby Internetu w poszukiwaniu zbiorów danych, a warunek ten spełniają te repozytoria, które na przykład mają zaimplementowany protokół OAI-PMH. Repozytoria takie powinny być również zarejestrowane w rozpoznawalnych na świecie nauki bazach danych repozytoriów, takich jak https://www.re3data.org a także zarejestrowane w OpenAIRE – Europejskiej Infrastrukturze Otwartego Dostępu.

Accessible – Dostępne

1. (Meta)dane są odnajdywalne poprzez otwarty, darmowy i uniwersalnie implementowany protokół
2. Działające linki
3. Protokół pozwala na procedurę uwierzytelniania i autoryzacji, jeśli jest to konieczne
4. Metadane są dostępne, nawet gdy dane nie są już dostępne

Interoperable – Interoperacyjne (przetwarzalne). Powinny być możliwe do przetworzenia dla ludzi i maszyn. Interoperacyjność oznacza możliwość wymiany i wykorzystania danych zarówno dla ludzi jak i maszyn.[8]

1. (Meta)dane są opisane z użyciem kontrolowanych słowników
2. Standardy metadanych
3. Format zapewniający odczyt maszynowy i możliwy do otwarcia przez ludzi

[8] M. Pawłowska, M. Wachowicz, Wprowadzenie do zarządzania danymi naukowymi, s.95

4. Porozumiewanie się z systemami – agregacja, indeksowanie
5. Powiązania do innych zasobów

Reusable - Możliwe do ponownego użycia.

1. (Meta)dane są bogato opisane, relewantność z daną dyscypliną
2. (Meta)danym jest przypisana licencja określająca warunki ponownego użycia
3. Otwarte formaty
4. Informacje dotyczące cytowania [9]

Repozytorium

Ze względu na wpływ naukowy istotne jest miejsce udostępnienia danych badawczych. Dane badawcze aby spełniały wymogi zasady FAIR należy udostępnić w dedykowanym repozytorium. Przy wyborze repozytorium należy wziąć pod uwagę poniższe kryteria.

Kryteria wyboru repozytorium

1. Zapewnienie trwałych i niepowtarzalnych identyfikatorów (PID) – repozytorium powinno nadawać trwały, unikalny identyfikator dla każdego zbioru danych (np. DOI).
2. Metadane – repozytorium powinno umożliwiać szczegółowe i dokładne opisanie zbioru. Metadane powinny być wyszukiwalne i możliwe do odczytu maszynowego. Informacje o danych powinny być publicznie dostępne nawet w przypadku danych niepublikowanych, chronionych, wycofanych lub usuniętych.
3. Dostęp do danych i warunki ich wykorzystania (licencje) - repozytorium powinno umożliwiać ponowne wykorzystanie danych na jasno opisanych warunkach (za: Pawłowska, Wprowadzenie do zarządzania danymi)
4. Ochrona – repozytorium powinno gwarantować bezpieczne długotrwałe przechowywanie danych[10]

[9] Więcej na stronie: https://www.go-fair.org/fair-principles/

[10] *Criteria for the selection of trustworthy repositories, Template from the Science Europe Practical Guide to the International Alignment of Research Data Management*, s. 2-3

Przy wyborze repozytorium zaleca się również branie pod uwagę certyfikacji np. Core Trust Seal[11].

W jaki sposób odnaleźć repozytorium danych badawczych?

Najlepszą praktyką ze względu na potencjalny wpływ naukowy jest zdeponowanie danych w repozytorium rozpoznawalnym w danej dyscyplinie. Repozytorium z określonej dyscypliny można odnaleźć korzystając z wyszukiwarki re3data. Instrukcja poniżej.

[11] Lista repozytoriów z ważnym certyfikatem: https://amt.coretrustseal.org/certificates

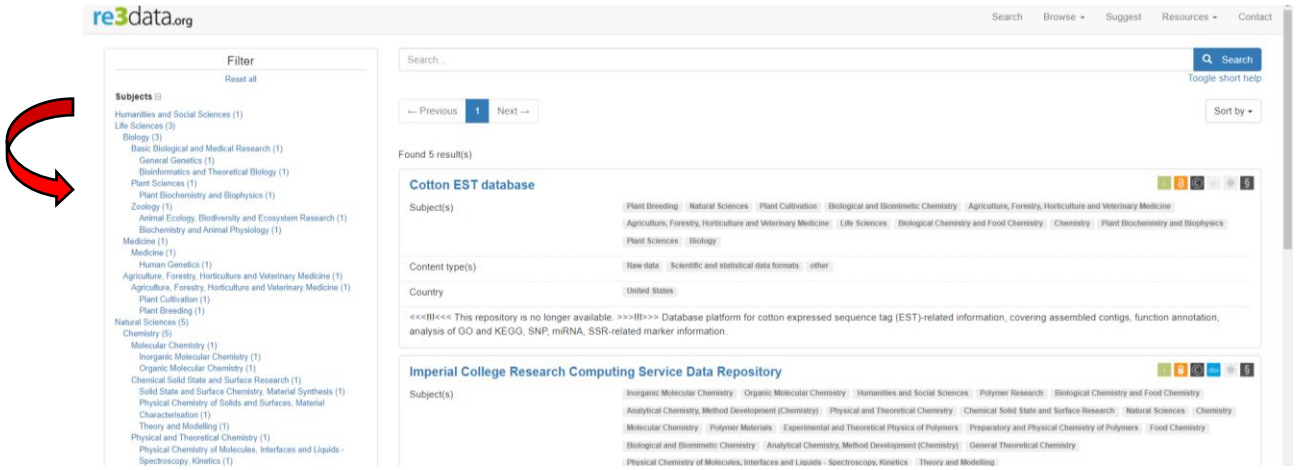
Wyszukiwarka re3data

Wpisujemy interesujące nas hasło przedmiotowe:

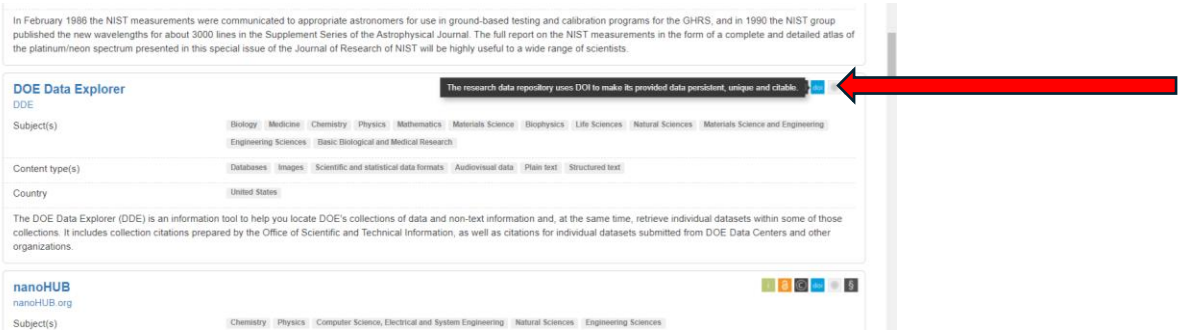
lub wybieramy z listy:

Następnie pojawi się koło dyscyplin na którym możemy wybrać dyscyplinę i zawęzić wyszukiwanie do interesującej nas tematyki:

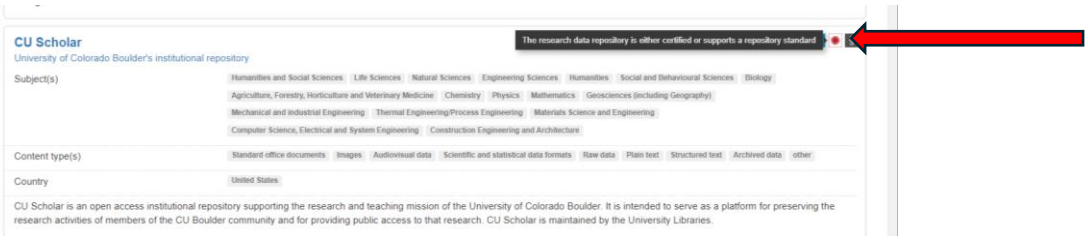
Można również wybrać dziedzinę korzystając z filtrów po lewej stronie ekranu:



Po wybraniu odpowiedniej opcji wyświetli nam się lista repozytoriów które posiadają w opisie to hasło przedmiotowe. Przy wyborze repozytorium należy zwrócić uwagę czy nadaje ono trwały unikalny identyfikator. Informuje o tym niebieska ikonka po prawej stronie ekranu:



Kolejną ważną cechą dobrego repozytorium jest certyfikat. O posiadaniu certyfikatu informuje ikonka z czerwonym kółkiem:



Udostępnianie danych badawczych

W tym rozdziale:

- Przygotowanie danych do udostępnienia
- Opis – metadane i plik readme
- Deponowanie danych badawczych w repozytorium RepOD

Przygotowanie danych do udostępnienia

1. **Sprawdzanie poprawności**

- Wartości odstające
- Brakujące wartości
- Kilka wartości w jednej komórce
- Wartości zarówno w wierszu jak i nagłówku kolumn
- Format daty, jednostki miary, itp
- Nagłówki kolumn
- Stosowane skróty[12]

W sprawdzaniu i czyszczeniu danych tabelarycznych mogą pomóc narzędzia tj SPSS czy STATA.

2. **Selekcja.**

Przed udostępnieniem danych badawczych należy zastanowić się nad ich selekcją. Nie wszystkie dane muszą być udostępnione. W procesie selekcji pomocne może być skorzystanie z pytań kontrolnych:

- Czy dane są niezbędne do walidacji opublikowanych wyników czy replikacji badań?
- Czy dane muszą być udostępnione/długotrwale archiwizowane w celu zapewnienia zgodności z polityką fundatora lub w celu spełnienia warunków określonych w umowie?
- Czy dane mają długoterminową wartość?
- Czy dane nadają się do ponownego wykorzystania?
- Czy potencjalny deponent jest właścicielem danych lub ma zgodę właścieli/osób badanych/współtwórców na udostępnianie danych?
- Czy dane są unikalne, czy mają wartość naukową lub historyczną?
- Czy planowany jest patent lub komercjalizacja?
- Czy objętość danych jest bardzo duża np. przekracza 100 GB? [13]

3. **Usunięcie danych osobowych** (danych umożliwiających identyfikacje osób) – anonimizacja.

Narzędzia do anonimizacji danych:

[12] Wickham, H. . (2014). Tidy Data. *Journal of Statistical Software*, *59*(10), https://doi.org/10.18637/jss.v059.i10 , s. 5-12.
[13] Więcej na: Data Selection and Appraisal Checklist (PDF)

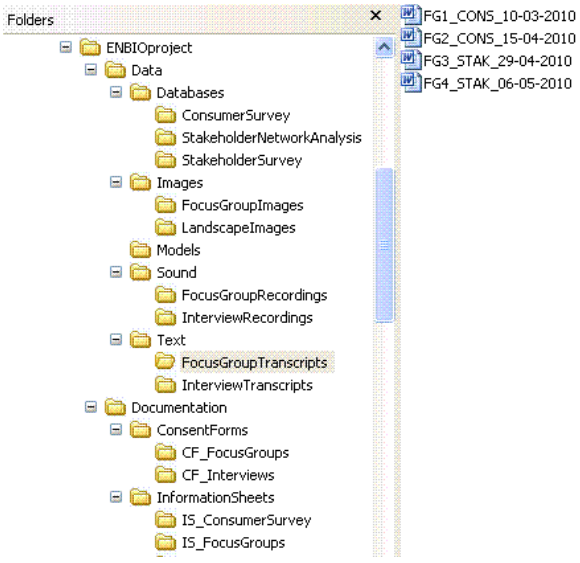
https://anonimizator.eadministracja.nask.pl/

https://amnesia.openaire.eu/index.html

4. **Nazewnictwo plików**
    - Nazwa pliku powinna być krótka, unikalna, powinna informować o zawartości i rodzaju pliku.
    - Daty powinno się zapisywać w formacie RRRR-MM-DD.
    - Nie należy używać spacji, kropek i znaków specjalnych (& lub ? lub !).
    - Do oddzielania elementów w nazwie pliku należy użyć myślników (-) lub podkreśleń (_).
    - W stosownych przypadkach w nazwach plików należy umieszczać oznaczenia wersji.

5. **Struktura folderów** – odpowiednie uporządkowanie plików w folderach ułatwia ich zlokalizowanie. Jest to szczególnie istotne przy współpracy z innymi. Jeśli kilka osób pracuje z danymi należy ustalić konwencję nazewniczą i sposób organizacji danych.

    Przykładowa struktura folderów
    W tym przykładzie pliki danych i dokumentacji są przechowywane w oddzielnych folderach. Pliki danych są dalej organizowane zgodnie z typem danych, a następnie zgodnie z działalnością badawczą. Pliki dokumentacji są również zorganizowane zgodnie z typem pliku dokumentacji i działalnością badawczą.



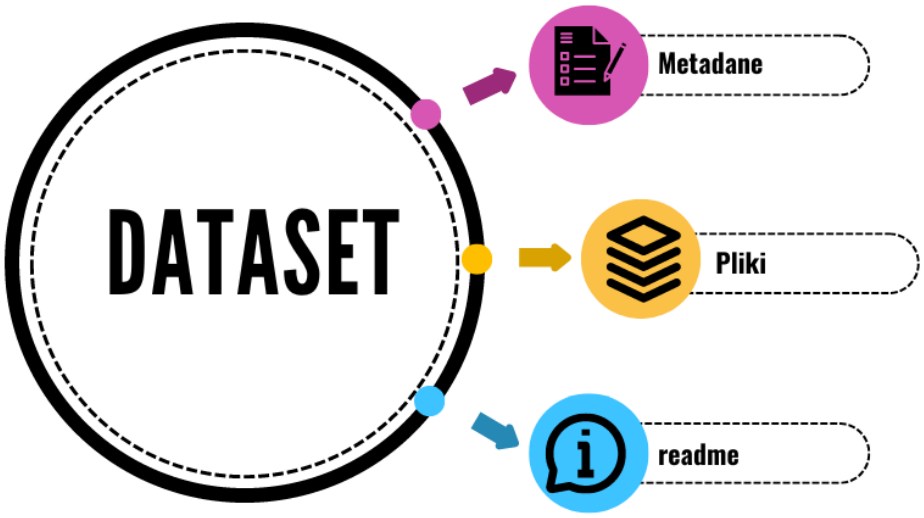
14

[14] Źródło: https://ukdataservice.ac.uk/

Opis

Dane i ich opis to nierozłączna całość. Sam plik z danymi może okazać się bezużyteczny, ponieważ będzie niezrozumiały dla innych.

Dataset – zbiór danych badawczych z odpowiednim opisem umożliwiającym zrozumienie i ponowne wykorzystanie danych.



Co powinien zawierać opis:

Opis powinien zawierać informacje o tym jak dane zostały zebrane, analizowane, przygotowane lub digitalizowane. Powinien być tworzony od samego początku zbierania danych i być na bieżąco aktualizowany. Opis pomaga nie tylko użytkownikom chcącym ponownie wykorzystać dane ale również twórcom rozpoznanie zawartości plików niezależnie od upływu czasu.

Cechy opisu wg standardu ALCOA:

1) Przypisany kontekst - źródło danych, kto i kiedy stworzył, jakie były modyfikacje
2) Czytelny – dokumentacja jest zrozumiała i możliwa do odczytania bez konieczności użycia specjalistycznego oprogramowania

3) Równoczesny - informacje o danych są zapisane w momencie generowania danych lub obserwacji;
4) Oryginalny - źródło zachowane w oryginalnej formie
5) Dokładny - bez błędów, wszelkie poprawki są udokumentowane [15]

Elementy opisu:

- **Cel, dla którego dane zostały utworzone**: Historia projektu, cele, założenia i hipotezy, badacze, fundatorzy i publikacje z danych.
- **Zawartość zbioru danych**: Rodzaj danych (np. wywiady, zdjęcia, odpowiedzi ankietowe), struktura plików danych, z liczbą przypadków, rekordów, plików i zmiennych, a także wszelkie relacje między takimi elementami.
- **Sposób gromadzenia danych**: Metodologia i protokoły gromadzenia danych, projekt doboru próby, struktura i reprezentacja próby, przepływy pracy, używane instrumenty, używany sprzęt i oprogramowanie, skala i rozdzielczość danych, stosowane metody digitalizacji lub transkrypcji, wykorzystywane wtórne źródła danych.
- **Kto zebrał dane, gdzie i kiedy**: Osoby zbierające dane, zasięg geograficzny zbierania danych, data zbierania danych (zasięg czasowy).
- **Sposób przetwarzania danych**: Narzędzia, instrumenty, procedury, sprzęt/oprogramowanie lub protokoły wykorzystane do edycji, czyszczenia, kodowania i klasyfikacji danych.
- **Wszelkie modyfikacje danych**: Anonimizacja, w przypadku badań szeregów czasowych lub badań podłużnych zmiany wprowadzone w metodologii, treści zmiennych, tekście pytań, oznaczaniu zmiennych, pomiarach lub doborze próby.
- **Procedury zapewnienia jakości**: Walidacja danych, sprawdzanie, weryfikacja, czyszczenie i inne przeprowadzone procedury zapewniania jakości, takie jak sprawdzanie sprzętu i błędów transkrypcji, procedury kalibracji, rozdzielczość przechwytywania danych i powtórzenia lub edycja, weryfikacja lub kontrola jakości materiałów.
- **W jaki sposób można uzyskać dostęp do danych**: Skąd dostępny jest zbiór danych, trwały identyfikator, warunki dostępu, warunki użytkowania, licencja, prawa autorskie, informacje o poufności danych, cytowanie.[16]

Standardy metadanych - aby dane były zrozumiałe dla algorytmów wyszukiwawczych

1. Metadane techniczne – generowane z urządzań badawczych i instrumentów
2. Metadane opisowe – mogą być tworzone ręcznie ale najlepiej korzystać z schematów i formatek dostępnych za pośrednictwem repozytoriów

[15] Źródło: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/e6r2-good-clinical-practice-integrated-addendum-ich-e6r1
[16] Źródło: https://ukdataservice.ac.uk/

Przykłady:

- Dublin core https://www.dublincore.org/specifications/dublin-core/
- DDI https://ddialliance.org/
- Crystallographic Information Framework https://www.iucr.org/resources/cif/spec
- Darwin Core https://www.tdwg.org/standards/dwc/
- Genome metadata https://www.bv-brc.org/docs/
- ISA-Tab - The Investigation/Study/Assay tab-delimited format http://isatab.sourceforge.net/docs/ISA-TAB_release-candidate-1_v1.0_24nov08.pdf
- INSPIRE https://knowledge-base.inspire.ec.europa.eu/index_en
- MIBBI (Minimum Information for Biological and Biomedical Investigations) https://rdamsc.bath.ac.uk/msc/m23 [17]

Słowniki kontrolowane i ontologie – pozwalają na jednoznaczny opis danych

- Słowniki i ontologie z dziedziny nauk humanistycznych: https://github.com/CLARIAH/awesome-humanities-ontologies
- Ontologie z dziedziny nauk biomedycznych: https://www.ebi.ac.uk/ols4/ https://bioportal.bioontology.org/
- Słowniki dziedzinowe: https://fairsharing.org/FAIRsharing.tnu8V7 https://vocabularies.cessda.eu/

Plik readme („przeczytaj mnie") – jest to plik który potencjalny użytkownik danych powinien przeczytać zanim otworzy właściwy plik z danymi badawczymi. Plik readme ma wyjaśniać innym użytkownikom co zawierają udostępnione pliki w zestawie danych, oraz pomóc w interpretacji i ponownej analizie zestawu danych. Powinien zawierać informacje dotyczące sposobu zebrania i przetworzenia danych oraz wszystkie informacje niezbędne do odczytu danych.

- Powinien być dołączony do plików zdeponowanych w repozytorium
- Powinien być zapisany w formacie tekstowym (txt)

[17] Więcej na stronie: https://rdamsc.bath.ac.uk/

Wzór zawartości pliku readme[18]:

1. Informacje podstawowe
   - Tytuł zbioru danych
   - Twórcy danych (Autorzy):
   - Instytucje wspierające

2. Opis badań i zbioru danych
   - Cel badan *[Opisz krótko cel badań, pytanie badawcze i kontekst, w jakim zbierano dane]*
   - Zakres danych:
     Zakres temporalny *[Przedział czasowy, np. „Dane zbierano od kwietnia 2020 do marca 2023”]*
     Zakres geograficzny *[Obszar, którego dotyczą dane, np. „Europa Środkowa; 45°N - 55°N, 5°E - 15°E”]*
   - Metoda zbierania danych
   - Aparatura

3. Zawartość zbioru danych
   - Lista plików np:
     data.csv – *Zawiera szczegółowe informacje o obserwacjach migracji.*
     metadata.json – *Plik z metadanymi w formacie JSON opisujący struktury danych.*
     readme.txt – *Dokumentacja zbioru danych.*
   - Struktura folderów *[Opisz strukturę folderów, jeśli dane są podzielone na grupy plików lub sekcje]*
   - Formaty plików *[Wymień formaty plików, np. CSV, JSON, TXT]*

4. Opis zmiennych
   - Zmienne i ich znaczenie np:
     species: *gatunek ptaka (string)*
     date: *data obserwacji (format: YYYY-MM-DD)*
     location: *lokalizacja obserwacji (string)*
   - Jednostki miary i kody *[Wyjaśnij jednostki miary i kody, np. brakujące wartości, kody dla zmiennych kategorialnych]*

[18] Na podstawie wzorów readme opracowanych przez Cornell University Data Services: https://data.research.cornell.edu/data-management/sharing/readme/ i Longwood Research Data Management https://datamanagement.hms.harvard.edu/collect-analyze/documentation-metadata/readme-files.

5. Metadane

- Standard metadanych *[Podaj nazwę standardu metadanych, jeśli użyto konkretnego standardu, np. Dublin Core]*
- Opis jakości danych:

[Informacje o walidacji danych, procedurach czyszczenia oraz ograniczeniach zbioru]

- Źródło danych:

[Podaj źródło danych, jeśli pochodzą z innych źródeł, np. publicznych baz danych]

6. Wymagania techniczne i sposób korzystania z danych

- Wymagania sprzętowe i programowe:

[Opisz, czy do korzystania z danych potrzebne są specjalne oprogramowanie lub zasoby sprzętowe]

- Instrukcje użytkowania:

[Wskazówki dotyczące ładowania danych, uruchamiania skryptów lub wstępnej obróbki danych]

7. Aktualizacje i historia wersji

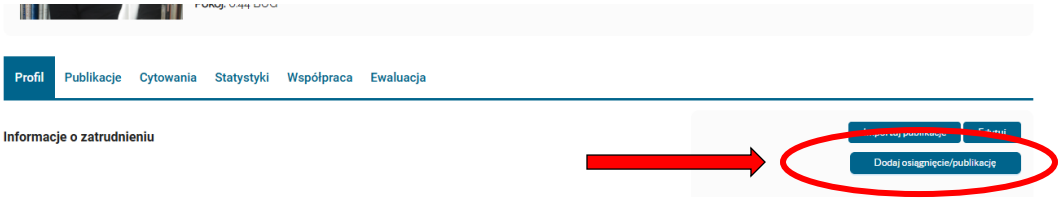
- Historia zmian np.:
  Wersja 1.0 (data): *[Opis wprowadzonej wersji]*
  Wersja 1.1 (data): *[Opis wprowadzonych zmian]*

Deponowanie danych badawczych w Bazie Wiedzy

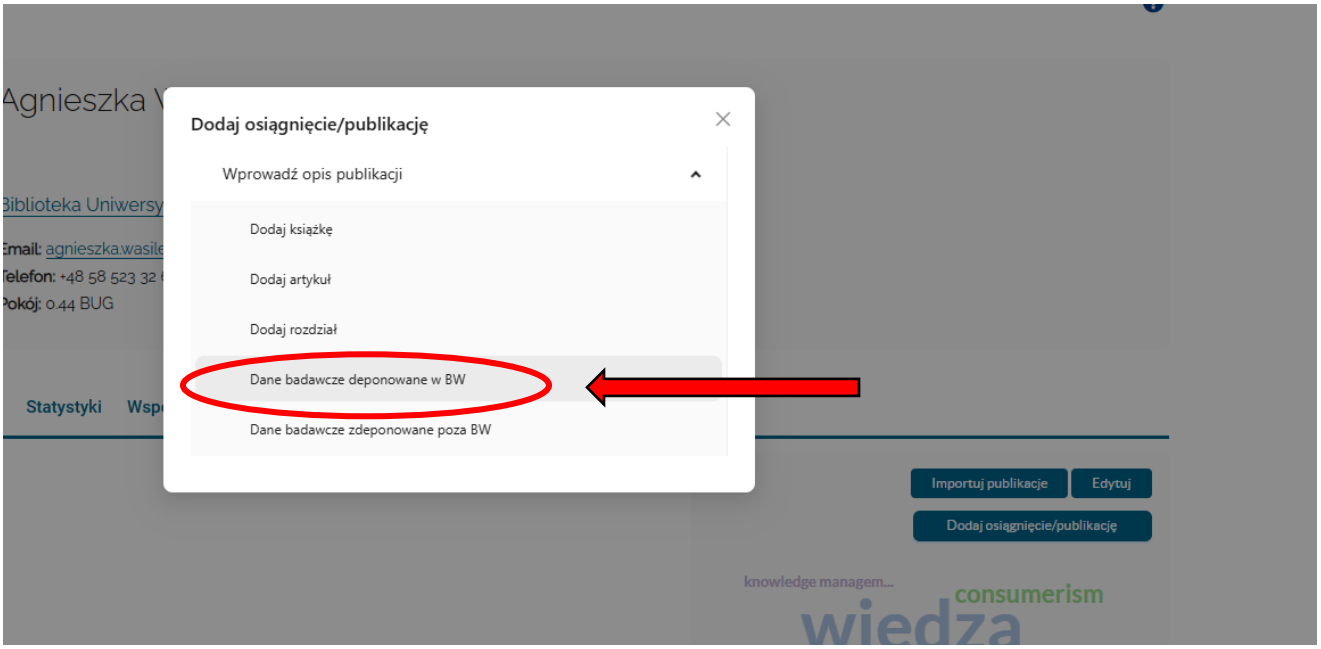
W repozytorium instytucjonalnym które jest częścią Bazy Wiedzy można zdeponować pliki nie przekraczające 50 MB. Zalecane jest deponowanie danych w formatach otwartych.

Dane badawcze można zdeponować samodzielnie, po zalogowaniu się do profilu autora.

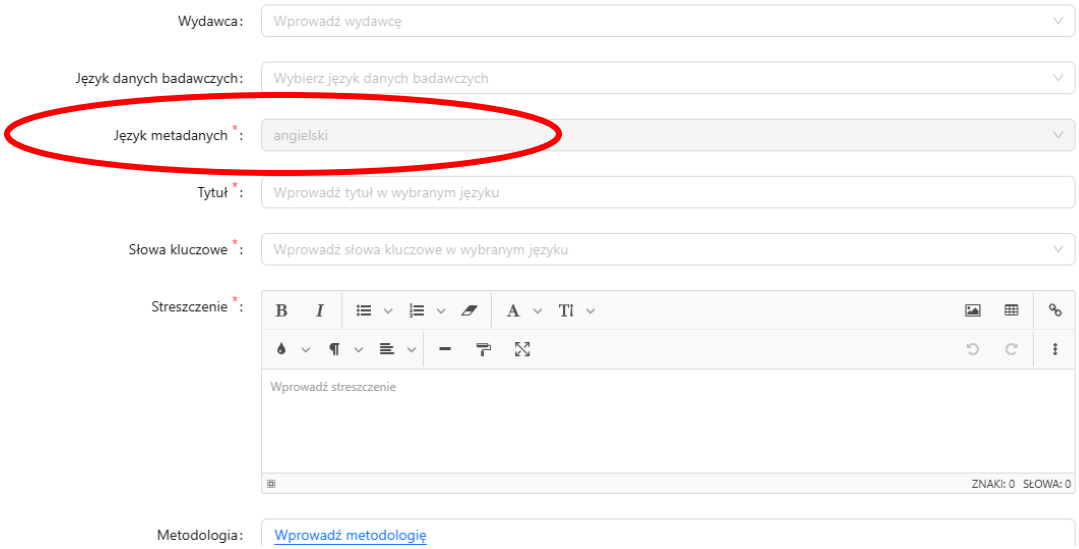
W profilu autora należy wybrać przycisk „Dodaj osiągnięcie/publikację”.



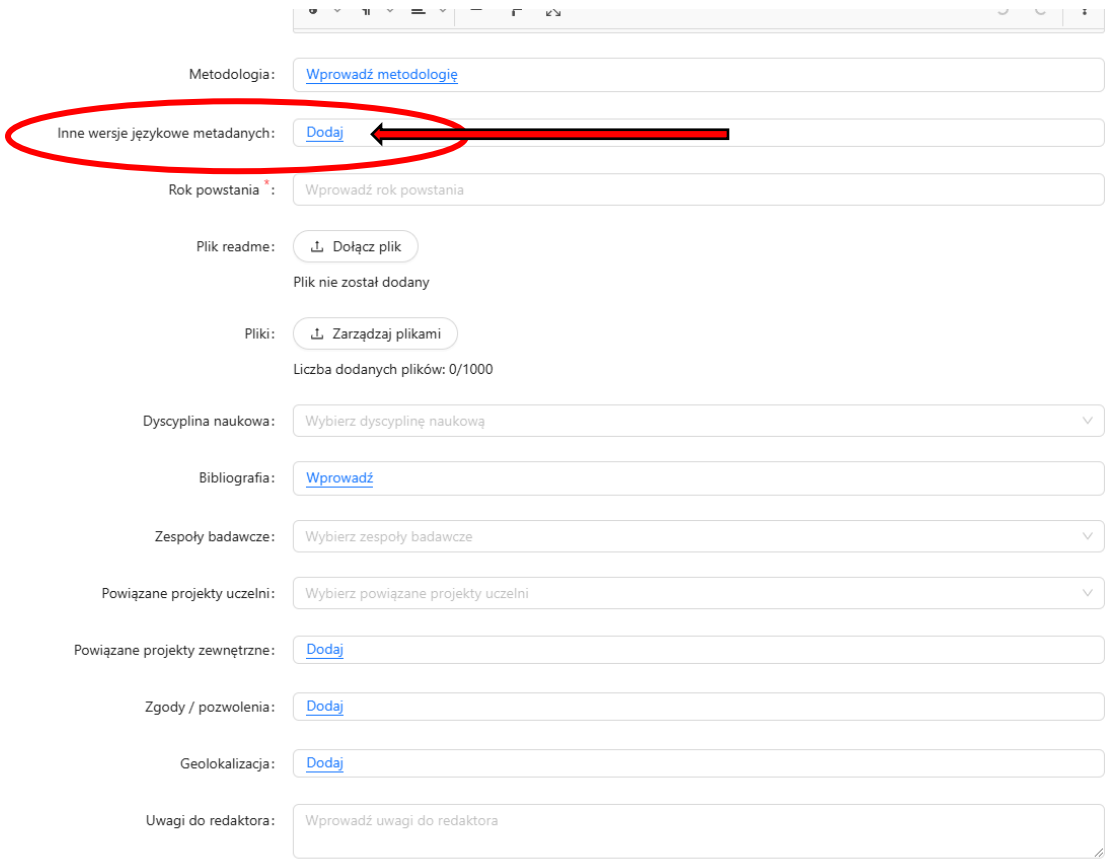
Następnie wybieramy z listy „Dane badawcze zdeponowane w BW”.



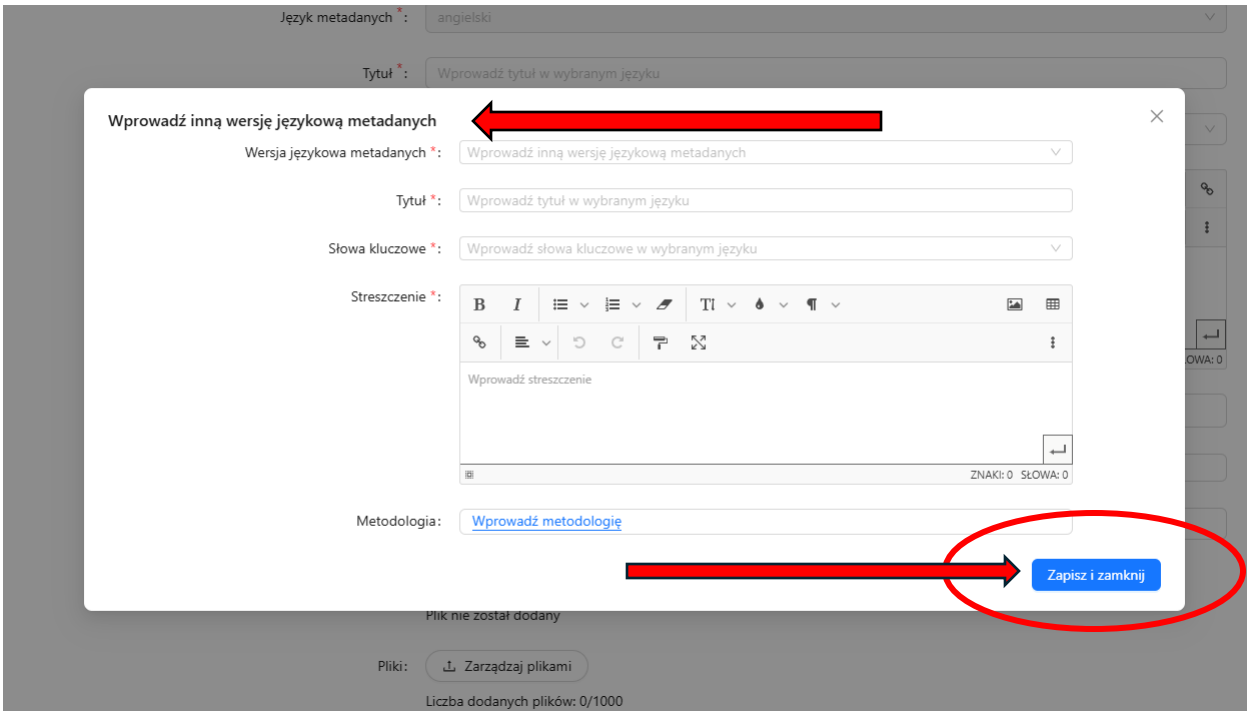
Wypełniamy pola metadanych. Preferowanym językiem metadanych jest angielski ze względu na indeksowanie w zewnętrznych bazach i międzynarodowych serwisach (tj. DataCite)



Jeśli chcemy wprowadzić inną wersję językową opisu wybieramy „Dodaj" w polu „Inne wersje językowe metadanych".

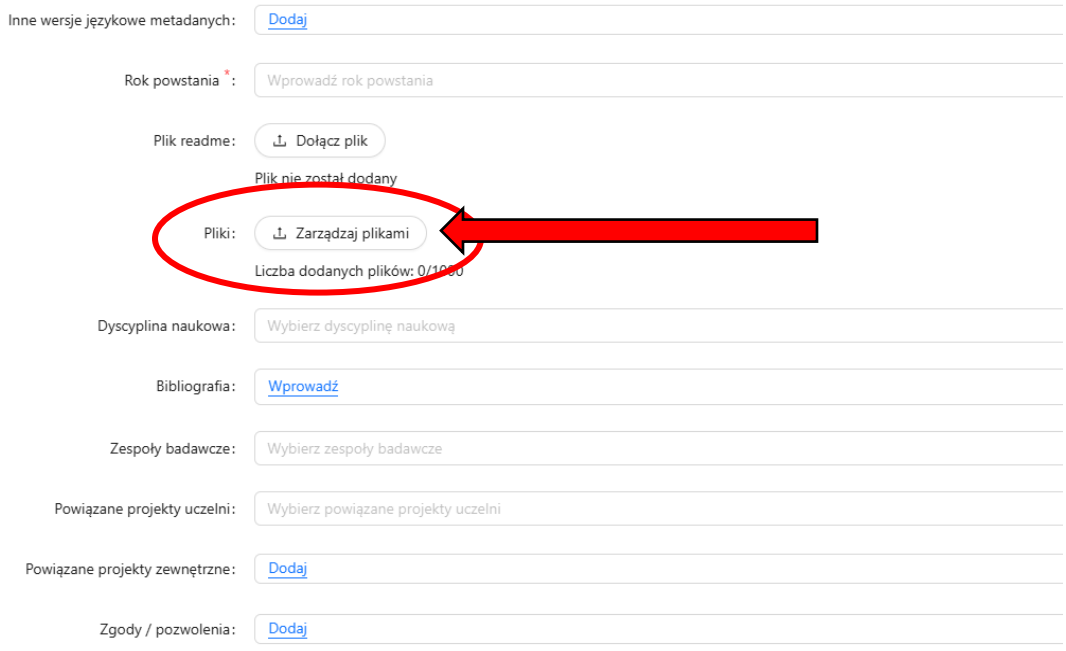


W nowym oknie wypełniamy pola innej wersji językowej opisu (np. w języku polskim) i klikamy przycisk „zapisz zmiany".

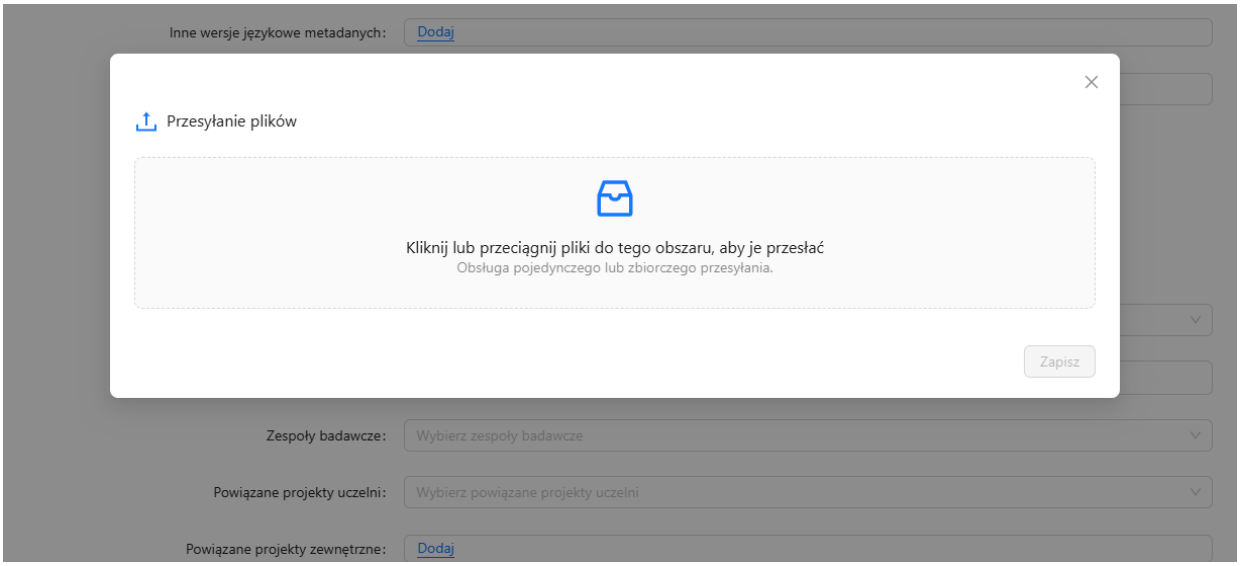


Wybieramy typ danych z listy podpowiedzi – do wyboru mamy pliki graficzne, filmy, pliki muzyczne, dane proste (np. dane tekstowe, arkusze kalkulacyjne, albo inne pliki w powszechnie znanych formatach), oprogramowanie lub inne.

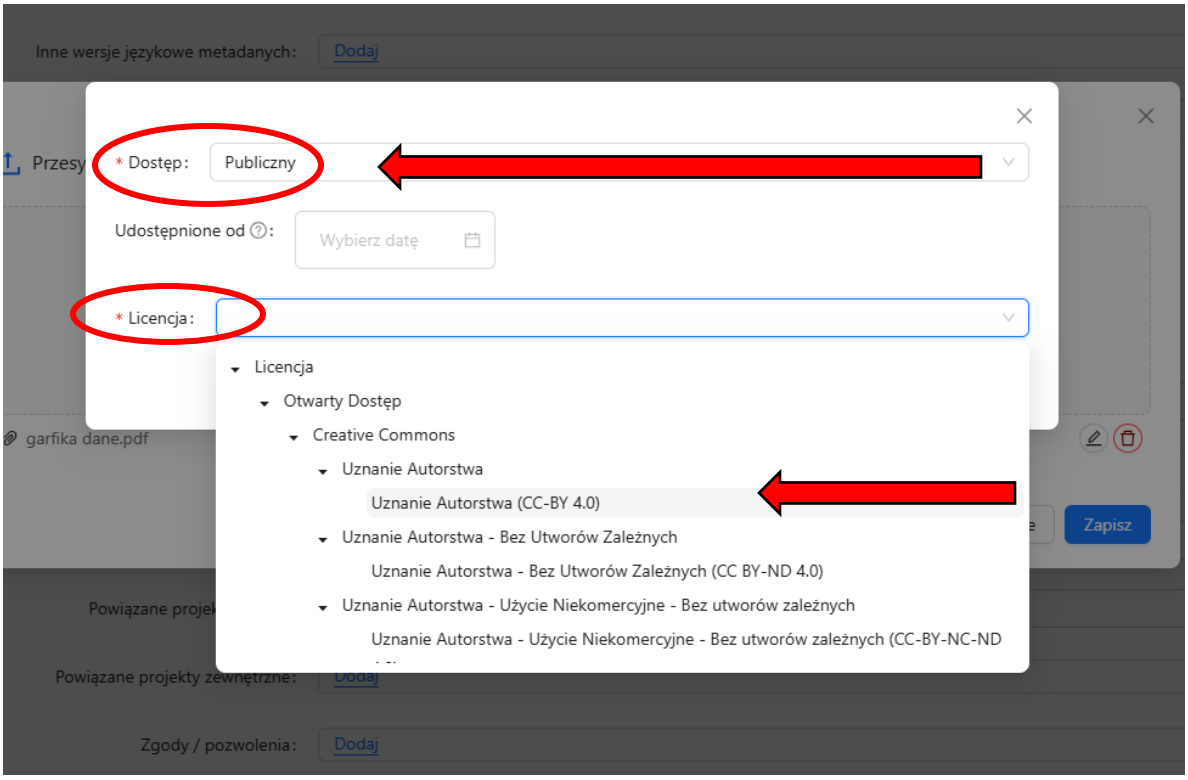
Aby dołączyć plik z danymi wybieramy przycisk „ Zarządzaj plikami”.



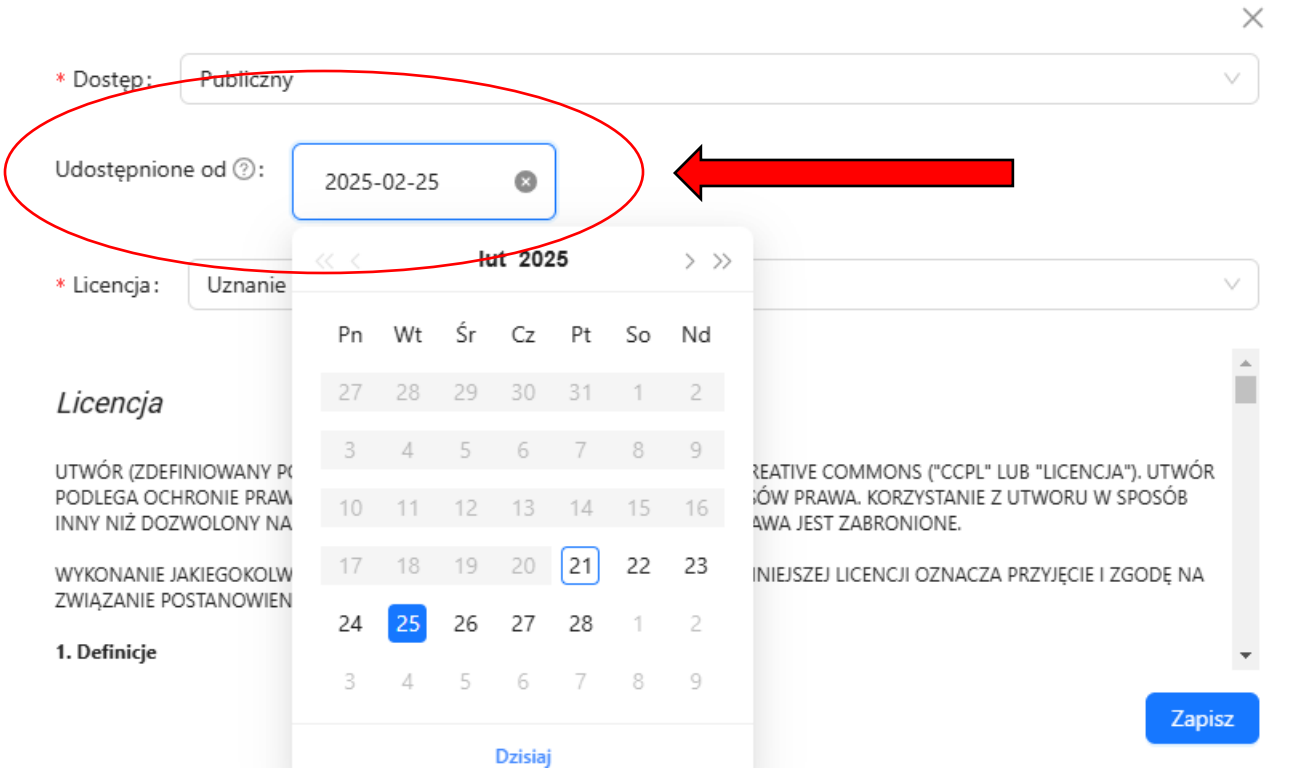
Następnie przeciągamy lub wybieramy pliki do przesłania.



Po załadowaniu pliku wybierz licencję i dostęp (chroniony tylko dla zalogowanych użytkowników lub publiczny) i kliknij „zapisz”.

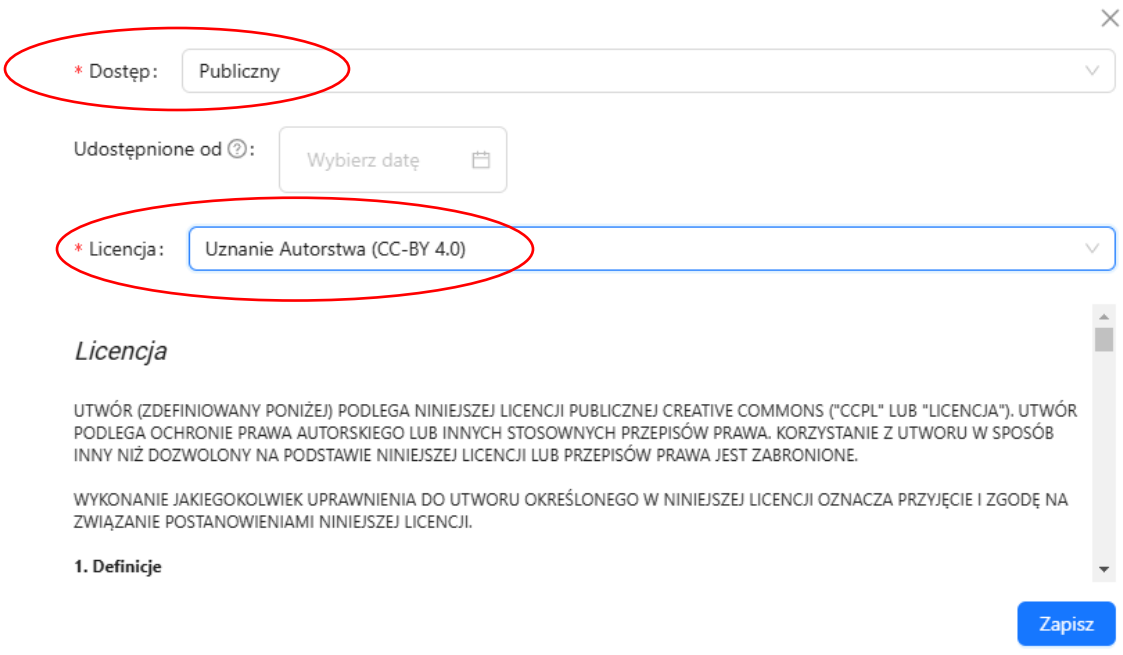


Jeśli chcemy ustawić embargo na dane wówczas wpisujemy datę od jakiej plik ma być dostępny publicznie. Do tego czasu plik będzie w dostępie zamkniętym.

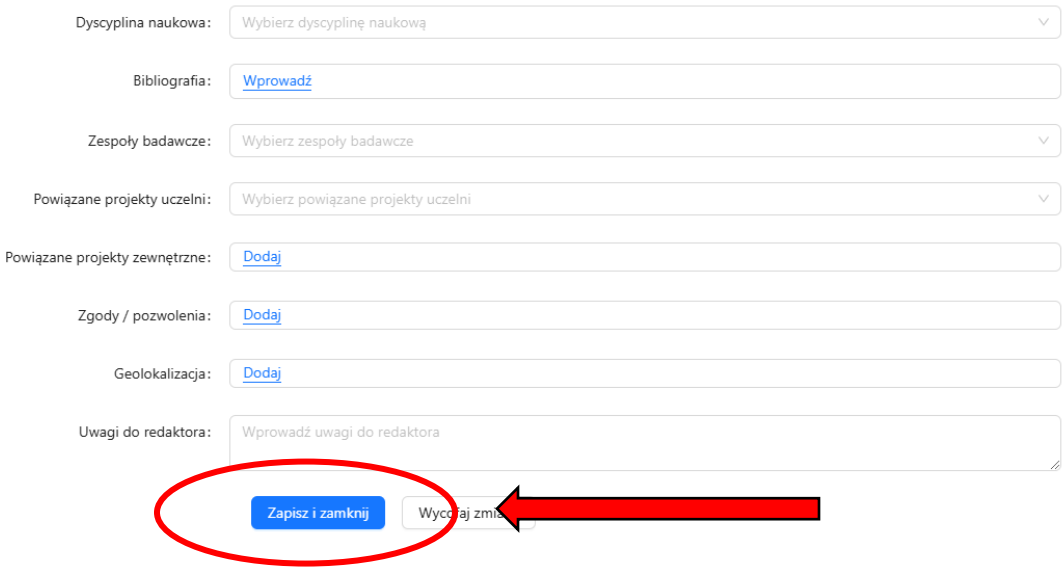


Do plików można dołączyć również plik readme w formacie tekstowym (txt.) zawierający informacje niezbędne do interpretacji i ponownego wykorzystania danych.

Po załadowaniu pliku readme z dysku wybieramy ikonkę edycji żeby nadać licencję i zdefiniować dostęp.



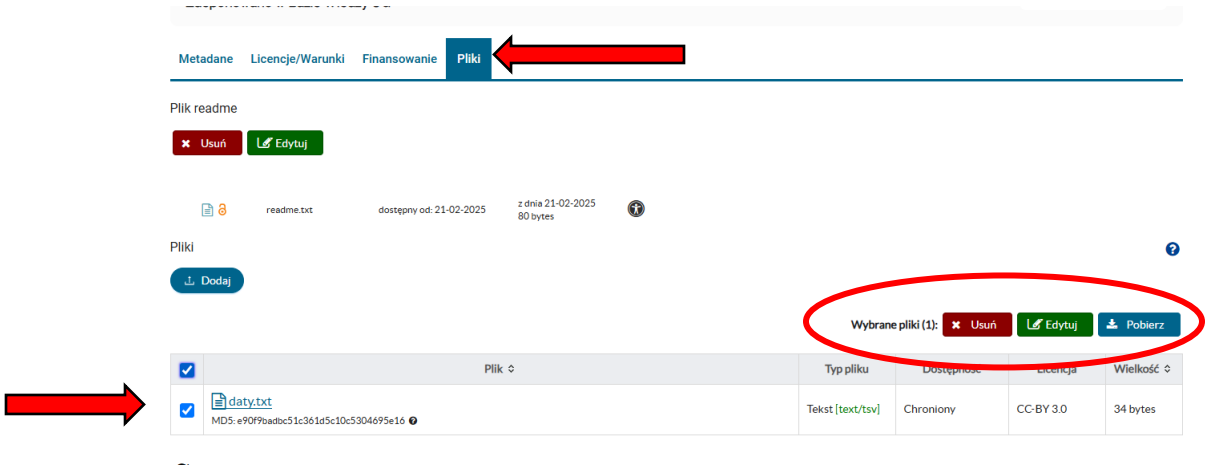
Na koniec wybieramy przycisk „Zapisz i zamknij”.



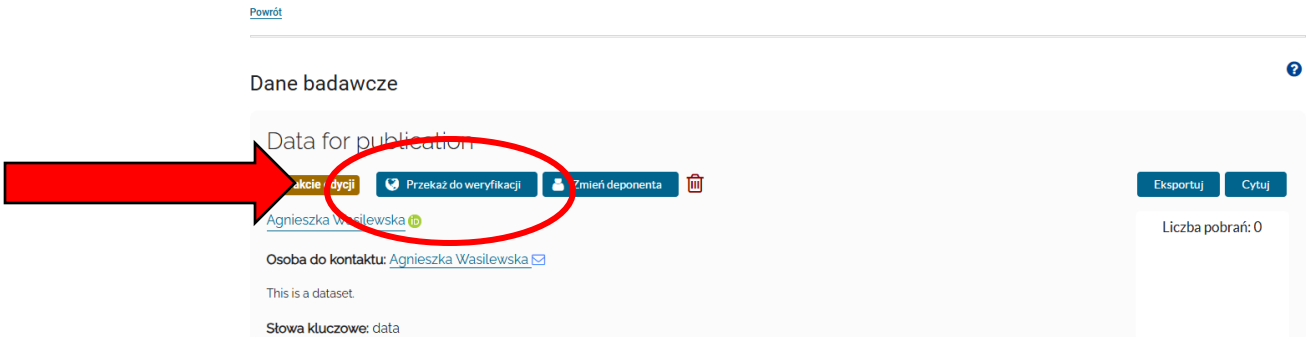
Tak zapisany rekord jest w wersji roboczej. Na tym etapie można nadal edytować dataset.

Aby edytować opis datasetu należy wybrać przycisk „edytuj” w zakładce metadane.

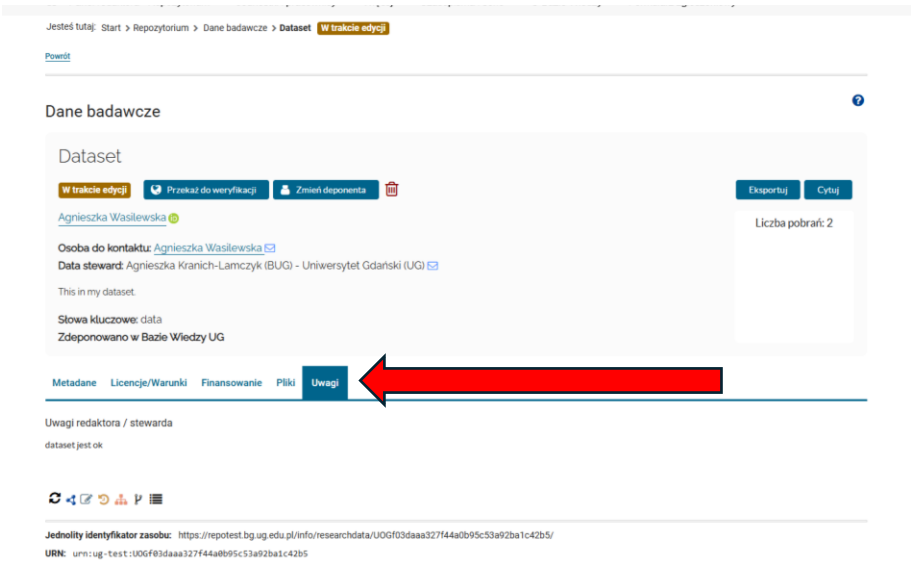
Aby edytować pliki należy wybrać zakładkę „pliki” a następnie zaznaczyć okienko przypliku który chcemy edytować. Wówczas uaktywnią się przyciski „edytu” oraz „usuń”.



Jeśli chcemy opublikować dane wybieramy przycisk „Przekaż do weryfikacji”. Zestaw danych jest weryfikowany przez redaktorów repozytorium a następnie publikowany.



Rekord w trakcie weryfikacji jest zablokowany do edycji. Wskazówki i uwagi data stewarda będą widoczne w zakładce uwagi.

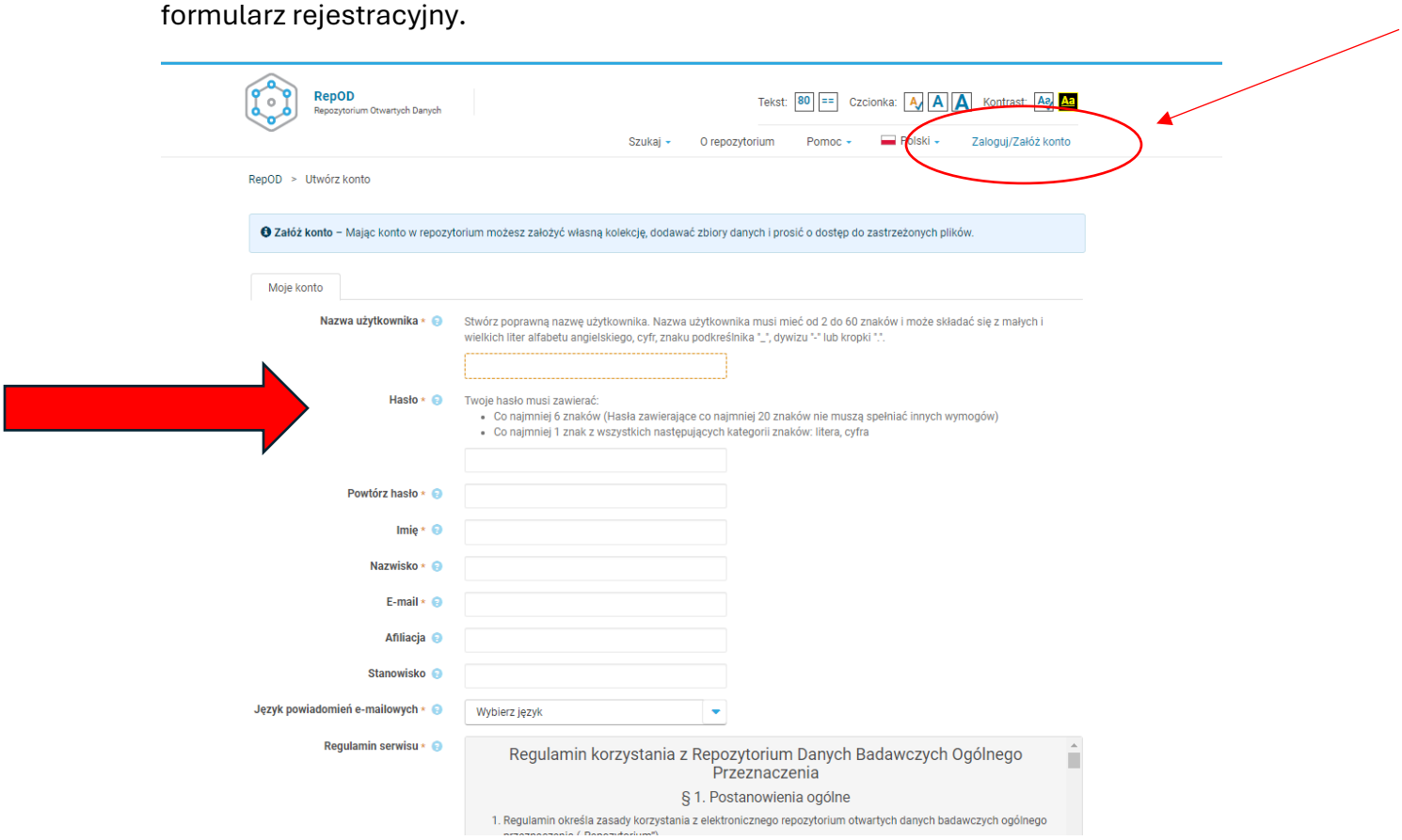


Po wprowadzeniu sugerowanych poprawek ponownie należy przekazać dataset do weryfikacji. Jeśli zbiór jest opisany prawidłowo redaktor zatwierdza go i publikuje.

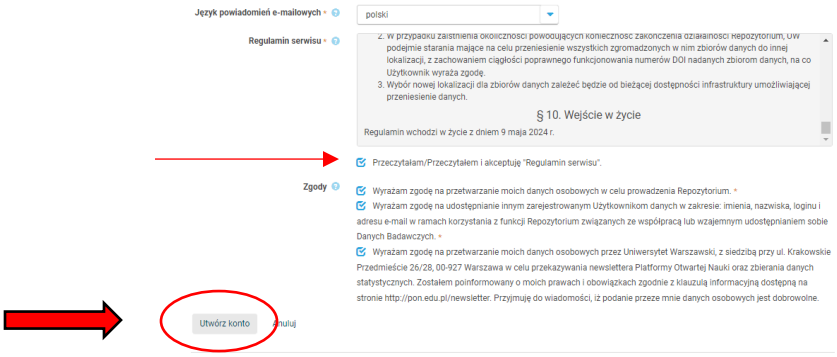
Deponowanie danych badawczych w repozytorium RepOD

Pliki powyżej 50 MB i większe datasety składające się z wielu plików można zdeponować w kolekcji instytucjonalnej Uniwersytetu Gdańskiego w repozytorium danych badawczych Repod. Deponowanie jest możliwe po utworzeniu konta i zalogowaniu się. Poniżej instrukcja:

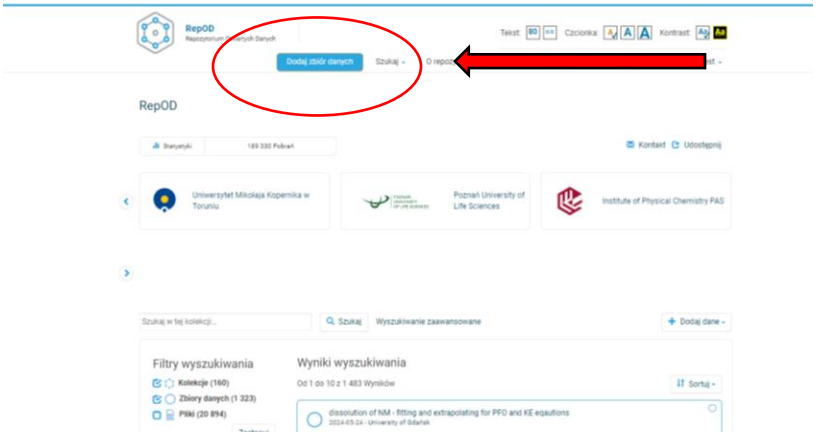
Na stronie Repod w prawym górnym rogu wybierz “załóż konto”. Następnie wypełnij formularz rejestracyjny.



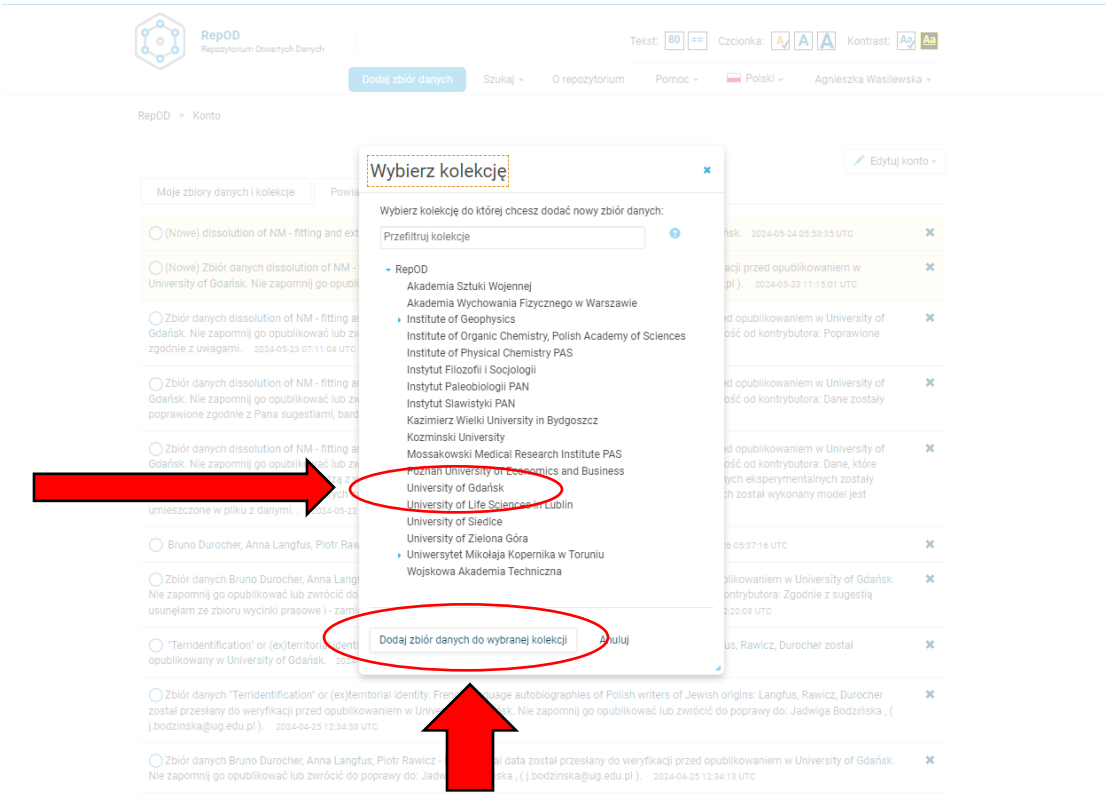
Zaakceptuj regulamin repozytorium i wybierz “utwórz konto”.



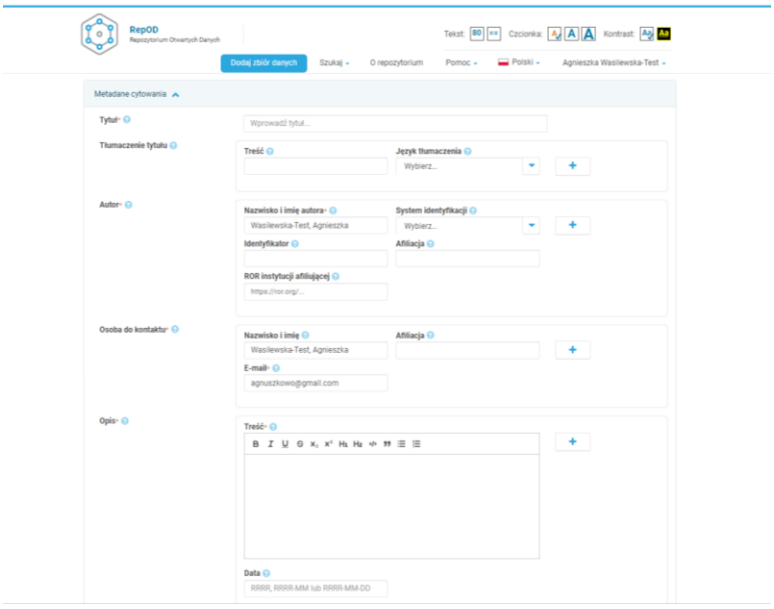
Po założeniu konta od razu możemy dodawać zbiory danych wybierając niebieski przycisk “dodaj zbiór danych” na górze strony:



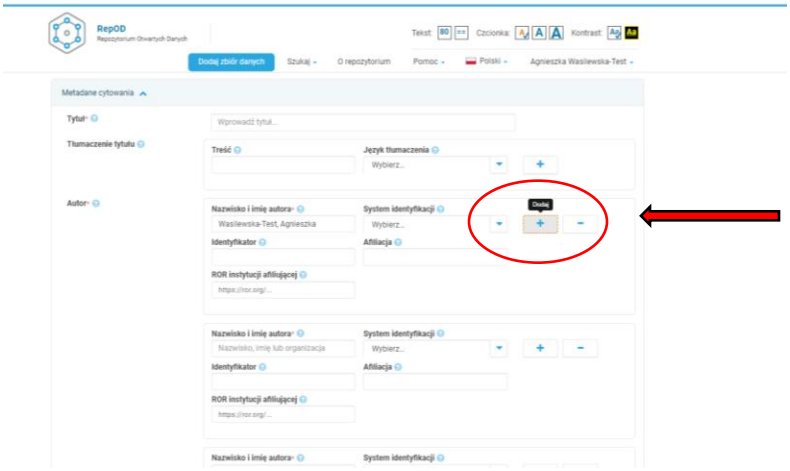
Wyświetli się okno “wybierz kolekcję”. Z wskazanej listy wybieramy kolekcję instytucjonalną UG klikając “University of Gdańsk”. Następnie potwierdzamy wybór klikając “dodaj zbiór danych do wybranej kolekcji”:



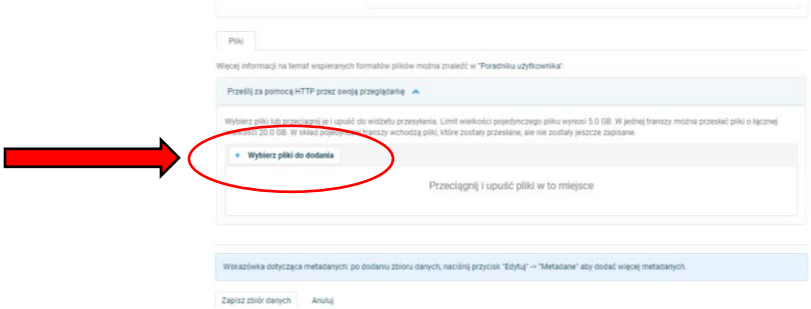
Następnie uzupełniamy formatkę metadanych:



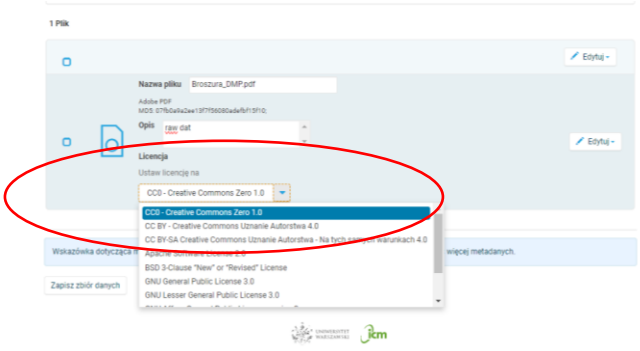
Kolejnych współautorów możemy dodawać klikając na plusik po prawej stornie:



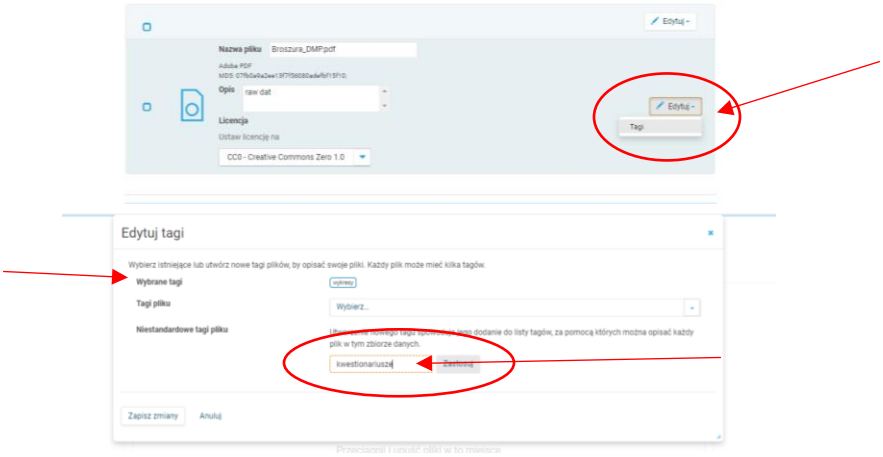
Po uzupełnieniu metadanych klikamy przycisk “wybierz pliki do dodania”



Po dodaniu pliku wybierz z listy odpowiednią licencję na jakiej chcesz udostępnić dane:

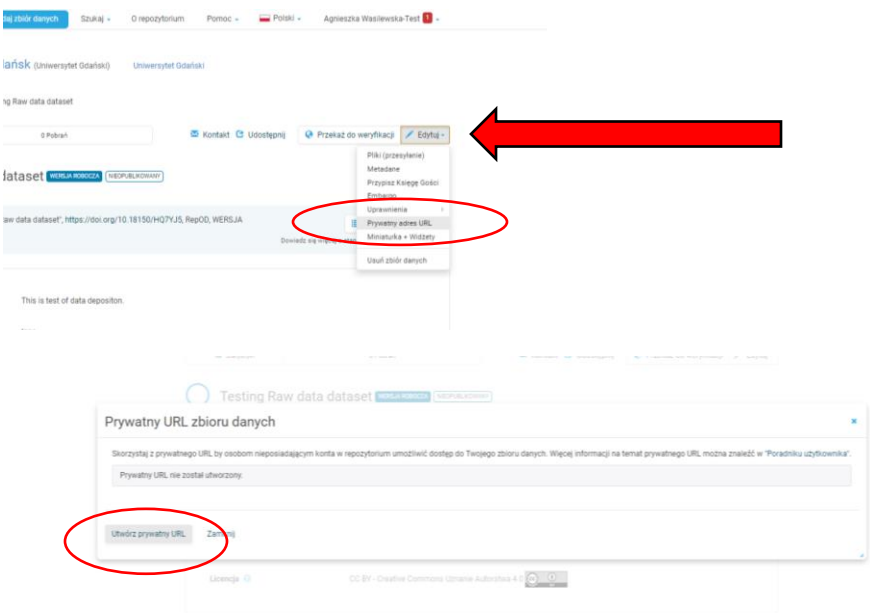


Po prawej stronie ekranu możesz wybrać przycisk “edytuj,” “tagi” aby dodać tagi opisujące zbiór:

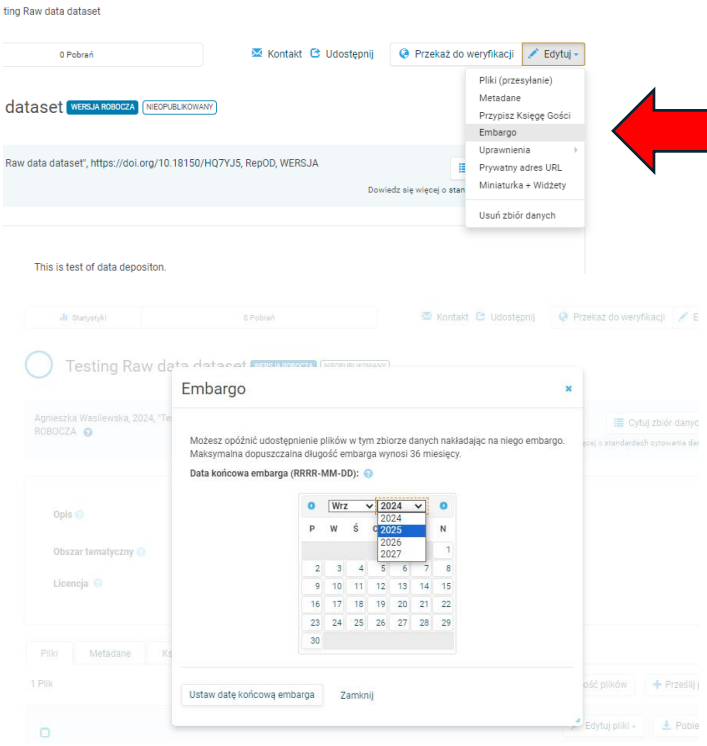


Na koniec wybieramy “zapisz zmiany” na dole strony:

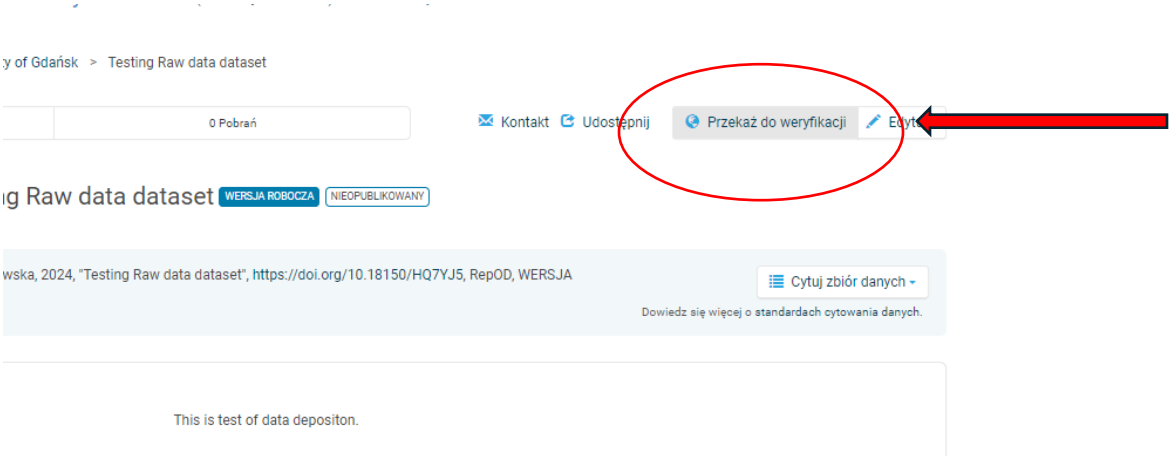
Utworzy nam się wersja robocza datasetu niewidoczna dla szerszej publiczności. Możemy wygenerować prywatny url (np dla recenzenta) wybierając na górze strony po prawej “edytuj”, “prywatny adres url”:



Można również zdefiniować embargo:



W tym momencie zbiór jeszcze możemy edytować, dodawać pliki i zmieniać metadane. Na koniec kiedy zbiór jest gotowy wybieramy po prawej stronie “przekaż do weryfikacji”:



Zbiór zostanie przekazany do weryfikacji administratorom kolekcji. W razie potrzeby administratorzy mogą odesłać informacje o konieczności wprowadzenia ewentualnych poprawek.

Ponowne użycie danych

W tym rozdziale:

- Cytowanie danych badawczych
- Jak odnaleźć dane badawcze
- Licencje

Cytowanie danych badawczych

Aby nie powtarzać kosztownych badań można skorzystać z już istniejących danych, udostępnionych na otwartych licencjach. Otwarte dane badawcze mogą być również bardzo cennym materiałem edukacyjnym. Podobnie jak w przypadku publikacji naukowych przy wykorzystaniu danych badawczych obowiązuje uznanie autorstwa.

Cytowanie powinno zawierać elementy:

- Autor/Twórca zbioru danych
- Data publikacji
- Tytuł
- Poziom przetwarzania danych (czy są to dane surowe czy przetworzone)
- Wersja
- Typ zasobu
- Wydawca
- Identyfikator
- Miejsce zdeponowania[19]

Wiele repozytoriów ma wbudowane wtyczki umożliwiające automatyczne wygenerowanie cytowań zgodnie z różnymi stylami.

Jak odnaleźć otwarte dane badawcze

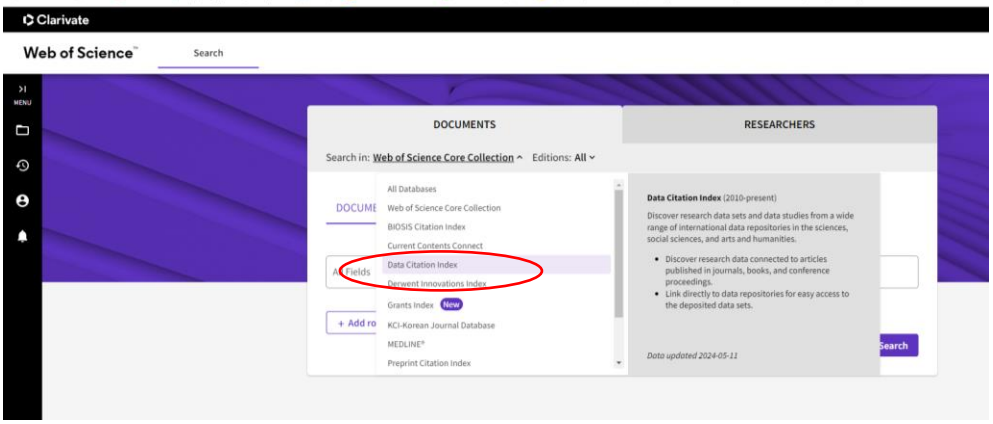
Aby odnaleźć dane badawcze można skorzystać z dedykowanych wyszukiwarek. Poniżej przedstawiamy instrukcję wyszukiwania danych badawczych w serwisach WOS data citation index, Google Dataset Search, DataCite oraz OpenAire explore.

[19] źródło: https://www.dcc.ac.uk/

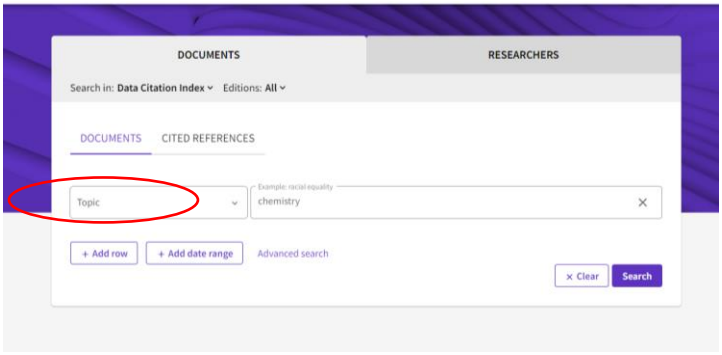
Wyszukiwanie danych badawczych

1. WOS data citation index:

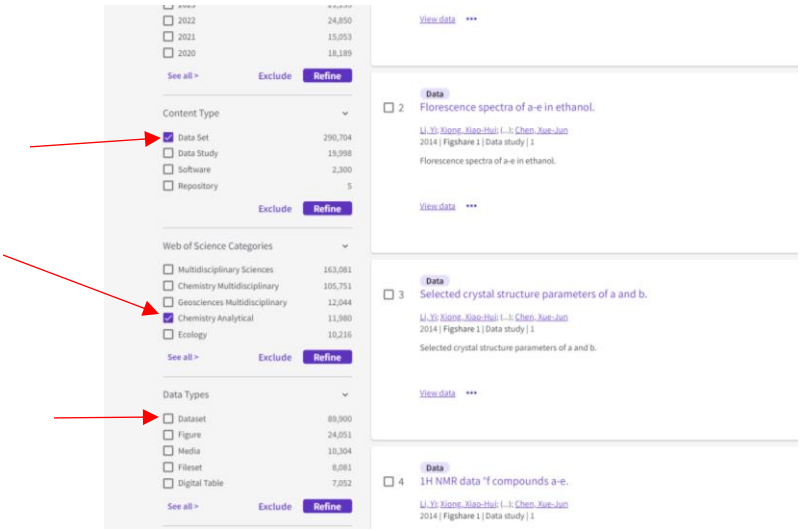
Na stronie WOS wybieramy opcję wyszukiwania (search in) Data Citation Index:



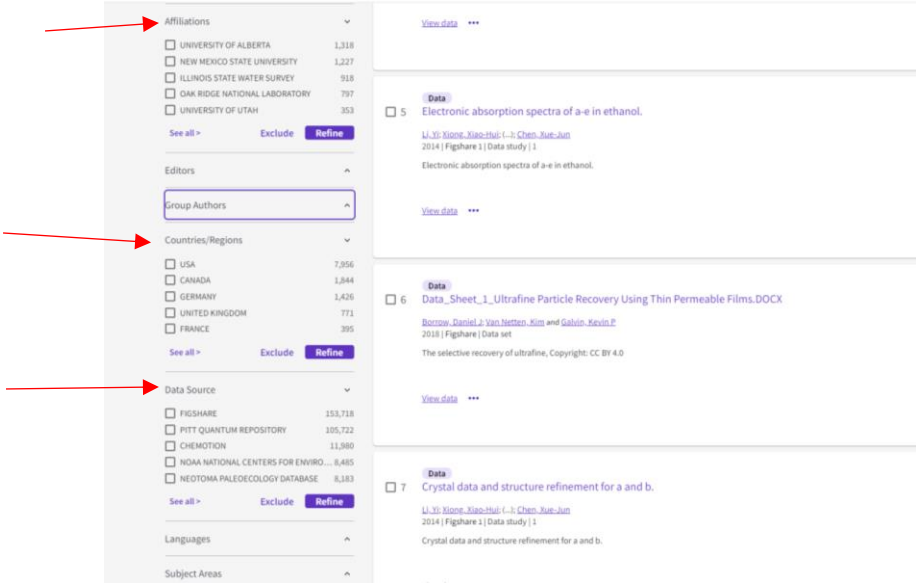
Możemy przeszukiwać bazę po tematyce:



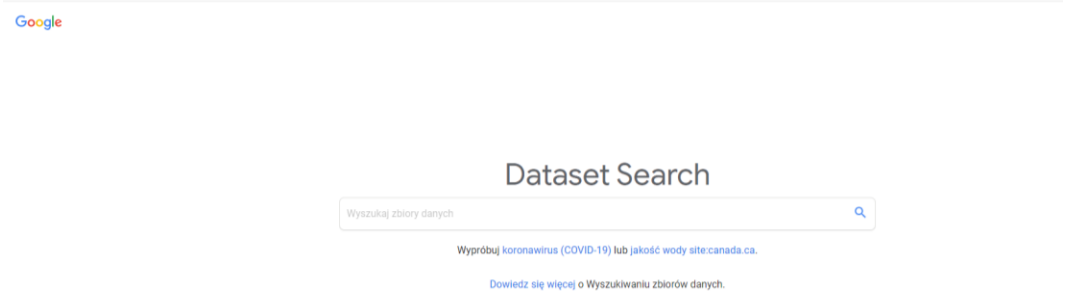
Następnie posługując się filtrami po lewej stronie ekranu możemy zawęzić wyszukiwanie do typu zbioru (data set) lub sprecyzować interesującą nas tematykę, możemy również wybrać konkretny rok publikacji:



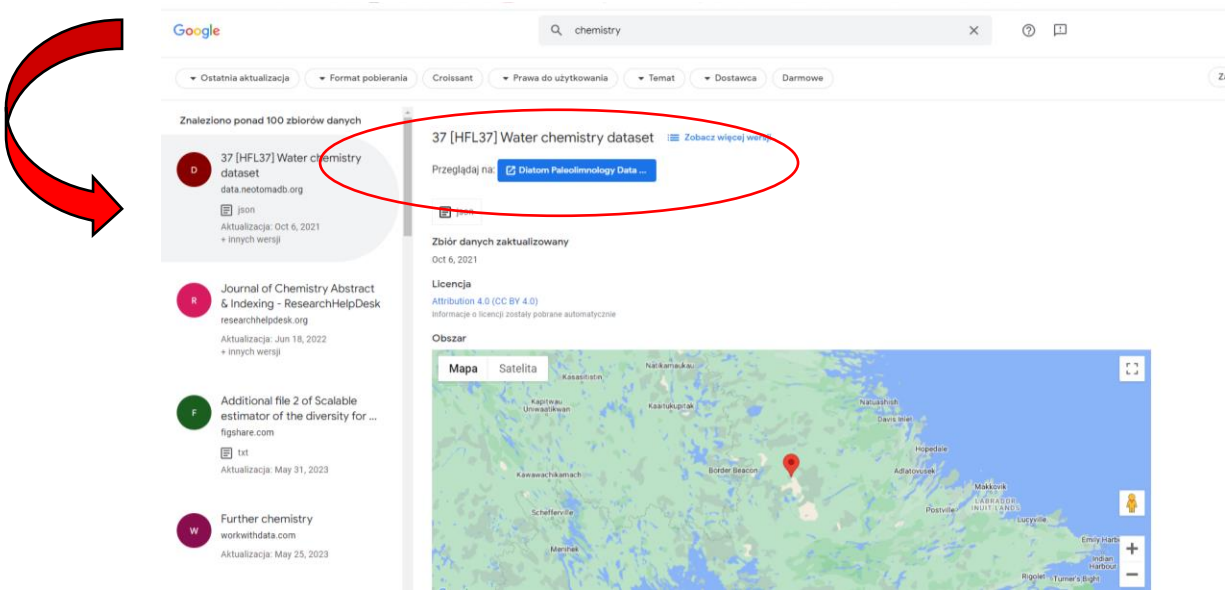
Filtry pozwalają ponadto określić m.in. afiliację, kraj czy źródło danych (repozytorium w którym zdeponowany jest zbiór):



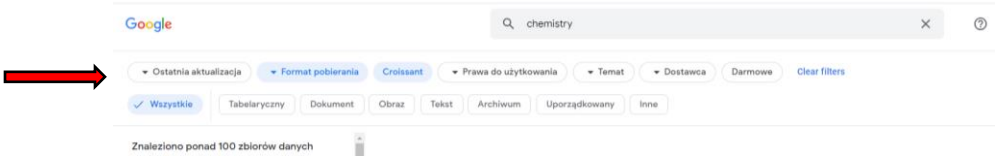
2. [Google Dataset Search]



Po wpisaniu hasła wyszukiwawczego rezultaty wyszukiwania wyświetlają się po lewej stronie ekranu a główną część po prawej zajmuje pierwszy wynik wyszukiwania – szczegółowe metadane pierwszego datasetu wraz z linkiem do źródła z którego pochodzi:

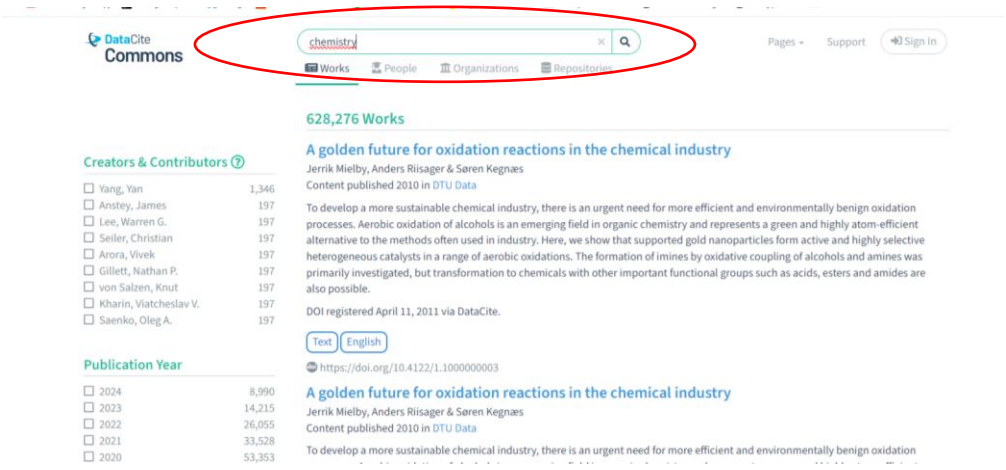


Rezultaty wyszukiwania można zawęzić używając filtrów na górze ekranu:



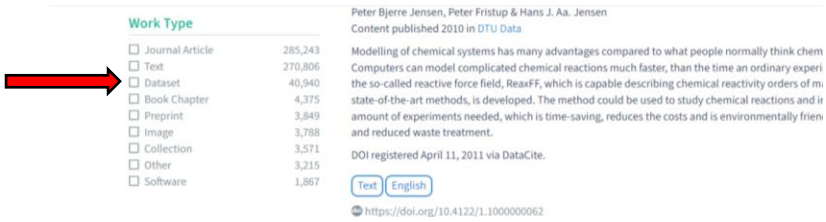
3. DataCite

W oknie wyszukiwawczym wpisujemy interesujące nas hasło:

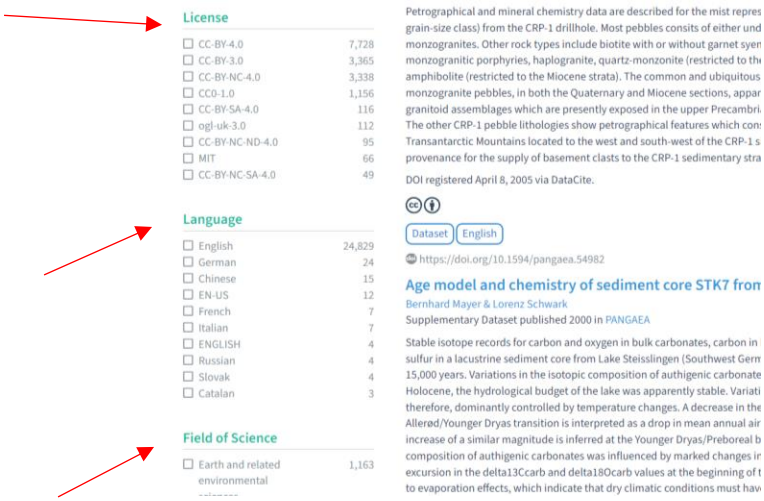


Wyniki wyszukiwania zawierają opisy danych badawczych z linkami do źródeł (do repozytorium w którym zdeponowany jest dany dataset).

Następnie posługując się filtrami po lewej stronie ekranu zawężamy wyszukiwanie do datasetów:

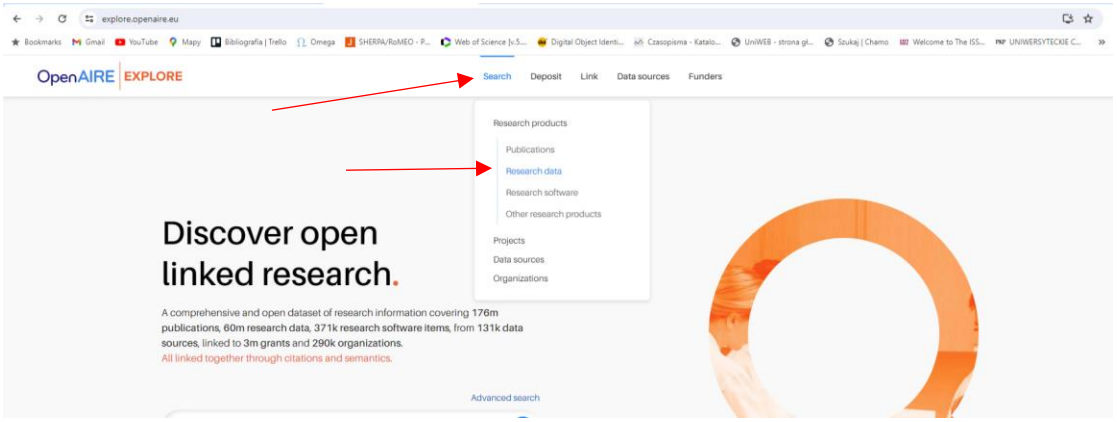


Możemy również doprecyzować licencję, język, dziedzinę nauki:

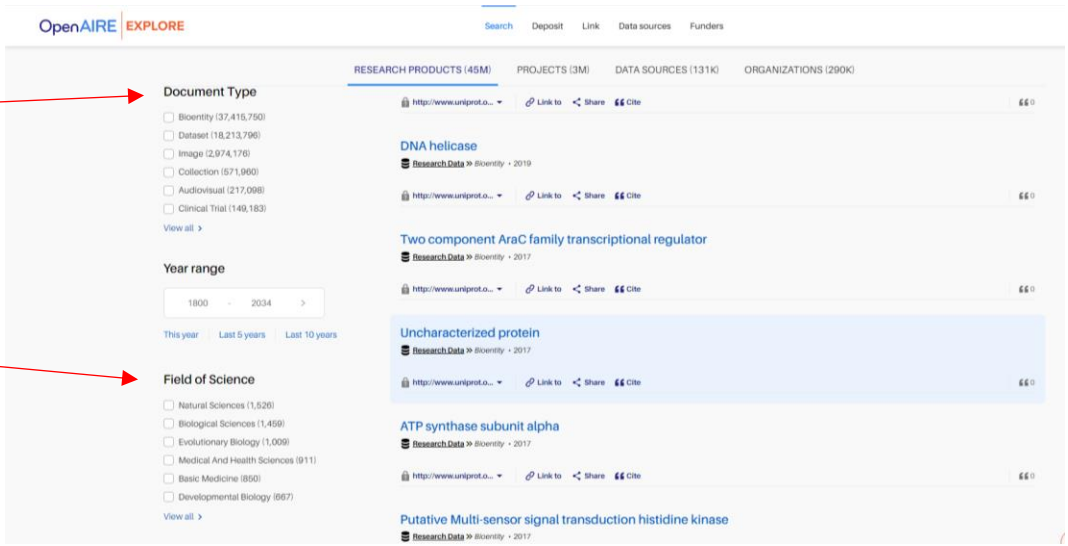


4. [OpenAire]

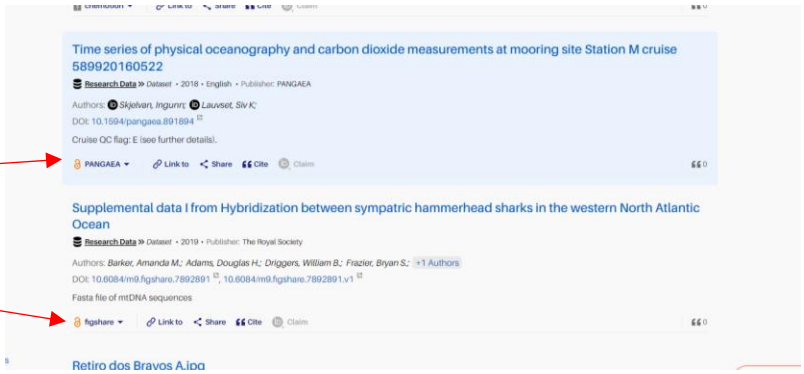
Na stronie OpenAire explore wybieramy z menu na górze strony „search” i „research data”



Następnie za pomocą filtrów po lewej stronie ekranu możemy zawężać wyniki naszego wyszukiwania o rodzaj danych czy dziedzinę nauki:



W opisie danych mamy informację o źródle i link bezpośrednio do repozytorium w którym zdeponowany jest dataset:



Licencje

W przypadku korzystania z istniejących danych należy sprawdzić na jakich zasadach można ich ponownie użyć.

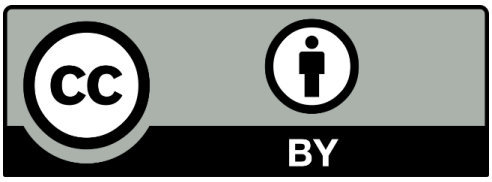
Licencje pozwalają określić w jakim zakresie można korzystać z udostępnionego utworu.

W przypadku wytworzonych danych to twórca nadaje określoną licencję danym, które udostępnia i tym samym decyduje co potencjalny odbiorca może zrobić z danymi.

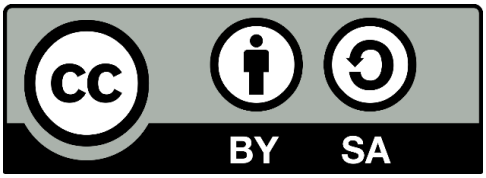
- **Wymóg NCN** – dane badawcze powinny być tam gdzie to możliwe udostępnione na licencji CC0. Od października 2023 w wyniku złagodzenia Polityki OA NCN pozwala na udostępnianie danych badawczych na licencji CCBY. [20]
- **Wymogi repozytoriów** – niektóre repozytoria wymagają przypisania deponowanym danym badawczym licencji CC0. Przed przesłaniem danych należy przeczytać regulamin i zapoznać się z warunkami.

Przykłady licencji:

Domena publiczna **-** zrzeczenie się praw autorskich.

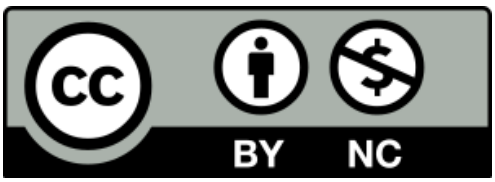


Uznanie autorstwa



Uznanie autorstwa, utwory zależne udostępniane na tych samych warunkach

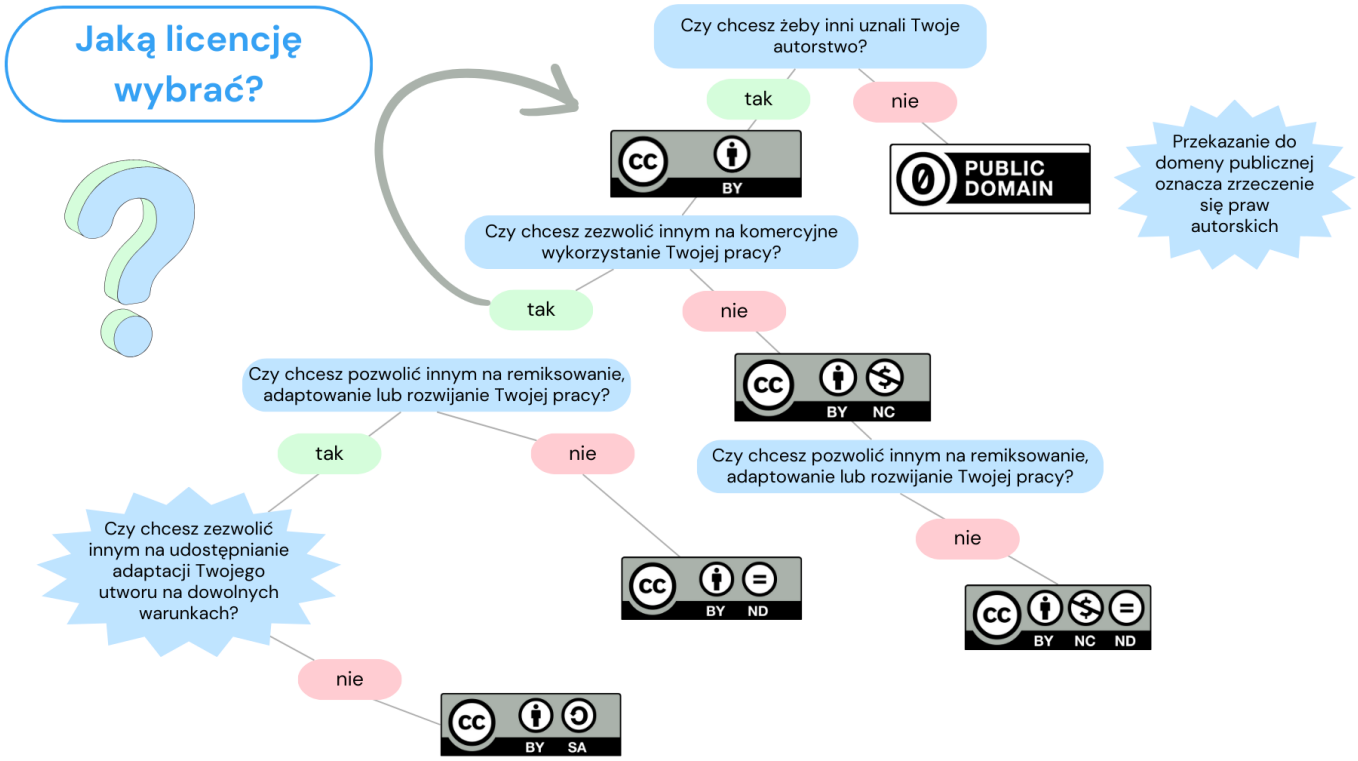
[20] Źródło: https://www.ncn.gov.pl/aktualnosci/2023-10-16-polityka-OA-zlagodzenie-przepisow



Uznanie autorstwa, użycie niekomercyjne

Uznanie autorstwa, użycie tylko niekomercyjne, nie wolno tworzyć utworów zależnych

Jak wybrać licencję:



Spis treści

Cel poradnika 1
Słownik pojęć 1
Zarządzanie danymi badawczymi 3
Dlaczego zarządzanie danymi przez naukowców realizujących badania naukowe jest ważne? ... 4
Kontrola jakości jako element zarządzania danymi badawczymi 5
Platforma Open Science Framework 6
Wybrane narzędzia wspomagające zarządzanie danymi 13
Planowanie w procesie zarządzania danymi 15
Cele planowania w procesie zarządzania danymi: 16
Podpowiedzi do poszczególnych punktów DMP: 17
Przechowywanie danych badawczych 31
Stosowanie zasady 3-2-1 32
Planowanie zapotrzebowania na przestrzeń dyskową w projekcie 32
Bezpieczeństwo danych w trakcie realizacji badań 32
Otwarte dane badawcze 34
Dane powiązane z publikacją 35
Deklaracje dotyczące dostępności danych (Data availability statements) 36
Zasada FAIR 36
Repozytorium 38
Udostępnianie danych badawczych 42
Przygotowanie danych do udostępnienia 43
Opis 45
Deponowanie danych badawczych w Bazie Wiedzy 50
Deponowanie danych badawczych w repozytorium RepOD 57
Ponowne użycie danych 63
Cytowanie danych badawczych 64
Wyszukiwanie danych badawczych 65
Licencje 71