



Zarządzanie Danymi Badawczymi

Archiwizacja i przechowywanie danych badawczych – wymagania, dobre praktyki, narzędzia

Szkolenie finansowane w ramach "Działań na rzecz promowania i zwiększenia świadomości w zakresie Otwartych Danych Badawczych", które stanowią element realizacji programu „Inicjatywa Doskonałości – Uczelnia Badawcza” (IDUB) w zakresie podniesienia poziomu jakości działalności naukowej uczelni w ramach Działania I.4 (Działania na rzecz zwiększenia liczby publikacji w prestiżowych czasopismach i wydawnictwach).



- Kluczowe aspekty archiwizacji i kopii zapasowych danych w kontekście RDM i DMP
- Korzyści płynące z archiwizacji danych badawczych dla przyszłych badań i rozwoju nauki
- Rola Data Stewarda
- Narzędzia na PG, źródła wiedzy



Przechowywanie i archiwizacja w cyklu zarządzania danymi badawczymi – zabezpieczenie danych

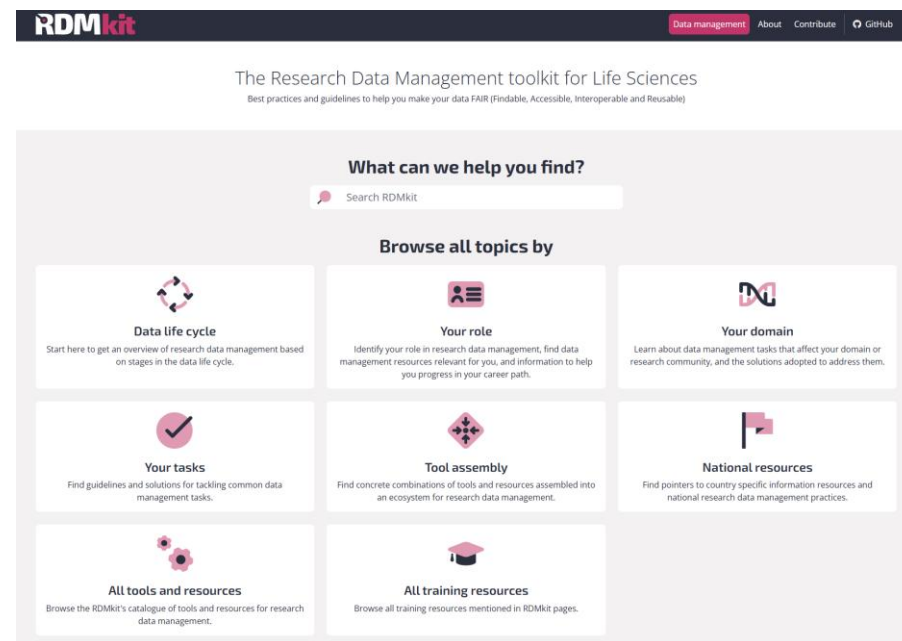
Archiwizacja/przechowywanie – proces długoterminowego przechowywania danych badawczych w celu ich udostępnienia i późniejszego wykorzystania

Kopia zapasowa – proces tworzenia tymczasowej kopii danych w celu regularnego zabezpieczenia przed utratą (np. spójności, kompletności, całości lub części danych)



Zapewnienie trwałości i dostępności danych

<https://rdmkit.elixir-europe.org>





Cykl zarządzania danymi badawczymi

<https://rdmkit.elixir-europe.org/planning>



Data Management Plan

*wszystkie aspekty związane z DMP
Planowanie ilości, jakości, zasobów, ról
odpowiedzialności, praw, licencji oraz
metod i miejsca zabezpieczenia i
archiwizacji danych*

<https://rdmkit.elixir-europe.org/collecting>



Pozyskiwanie danych

*DMP punkt 1 i 2
Pozyskiwanie danych, dbanie o jakość i
spójność danych, "pilnowanie" formatów,
zbieranie/budowanie metadanych*

dostępność danych w ramach
zespołu
! kopia zapasowa

<https://rdmkit.elixir-europe.org/processing>

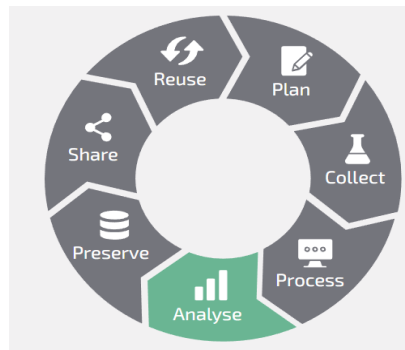


Przygotowanie danych do analizy

*DMP punkt 1 i 2, 4
Porządkowanie, konwertowanie,
anonimizacja, filtrowanie,
ujednolicanie, weryfikacja*

dostępność danych w ramach
zespołu
! kopia zapasowa

<https://rdmkit.elixir-europe.org/analysing>



Analiza danych

*Dobór narzędzi, infrastruktury, praca z
danymi w celu osiągnięcia wyników
badań,*

dostępność danych w ramach
zespołu
! kopia zapasowa



Przestrzeń na dane wymagana jest już na etapie zbierania i generowania danych!!!

O przechowywaniu danych pomyśl już na etapie **planowania**:

- objętość danych
- zasady dostępu
- potrzeba współdzielenia z członkami zespołu
- rozdzielenie i zabezpieczenie danych surowych/źródłowych
- przestrzeń i zasoby na przetwarzanie i analizowanie
 - uwzględnij narzędzia i usługi do przetwarzania danych
 - uwzględnij wydajność dysków, komputerów, serwerów
- wersjonowanie, kontrola spójności, śledzenie zmian

Następnie uwzględnij:

- kopia zapasowa
- archiwizacja i udostępnianie



Archiwizacja danych:

- długoterminowe przechowywanie danych badawczych po zakończeniu projektu
- cele: umożliwienie weryfikacji wyników badań, ponownego wykorzystania danych, spełnienie wymogów prawnych i instytucjonalnych
- dane archiwizowane są zazwyczaj niezmiennie, dobrze udokumentowane, przechowywane w formatach trwałych (np. CSV, PDF/A)

Kopia zapasowa:

- tymczasowa kopia danych tworzona regularnie w celu zabezpieczenia przed utratą w wyniku awarii sprzętu, błędu użytkownika, ataku hakerskiego
- backupy mają charakter operacyjny, a nie archiwalny
- przechowywane lokalnie lub w chmurze, często nadpisywane, oferowane "automatycznie" przez narzędzia do archiwizacji, przechowywania i udostępniania

Główne różnice:

- cel: trwałe przechowywanie vs. odzyskiwanie w przypadku awarii
- czas przechowywania: lata vs. dni/tygodnie
- typ danych: dane zakończonych projektów vs. aktywne dane robocze lub kopie archiwizowanych danych

Jakie dane?

Kluczowa jest analiza jakie rodzaje danych powinny zostać zarchiwizowane:

- dane źródłowe: dane zbierane, tworzone lub przechowywane w innych miejscach, które zostały wykorzystane w badaniach.
- zestawy danych: dane wyodrębnione lub pochodne z danych źródłowych.
- dane referencyjne: dane przekształcone w celu pogłębienia analiz lub wyciągnięcia wniosków/osiągnięcia wyników
- „materiały uzupełniające” w badaniach (zdigitalizowane notatki ręczne, dokumentacje itp.)
- dane wrażliwe (anonimizacja)
- dokumentacja

Jak długo?

Kluczowa jest określenie jak długo dane powinny być archiwizowane:

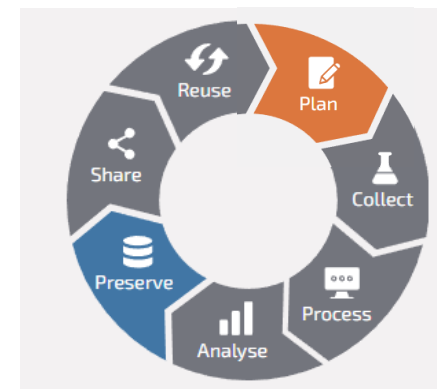
- wymagania grantodawców (NCN)
- przepisy prawne lub umowne
- wymagania instytucji
- wymagania organizacji partnerskich



Koszty?

koszty związane z archiwizacją

- przygotowanie danych do archiwizacji
 - Koszty pracy osób przygotowujących dane (opracowanie, ujednolicanie formatów, usuwanie zbędnych plików, konwersja do formatów zalecanych do długoterminowego przechowywania).
 - Możliwe koszty licencji na oprogramowanie do konwersji, kompresji lub walidacji danych.
- Porządkowanie danych
 - Prace polegające na uporządkowaniu struktury folderów, spójności nazw plików, wersjonowaniu.
 - Ewentualne koszty narzędzi do automatyzacji tych procesów.
- Dokumentowanie
 - Koszty pracy przy tworzeniu dokumentacji (metadane, README, opis metod, słowniki zmiennych).
 - konsultacje
- Przestrzeń dyskowa
 - Przestrzeń na archiwizację (opłaty za długoterminowe przechowywanie danych)
 - Przestrzeń na kopie zapasowe (opłaty za przestrzeń dyskową uwzględniającą miejsce na kopie zapasowe)
- Długoterminowy dostęp
 - Utrzymanie dostępności danych przez wymagany okres (np. 5–10 lat).
 - Możliwe koszty aktualizacji formatów plików w przyszłości, aby zapewnić ich dostępność
 - Koszty administracyjne związane z utrzymaniem repozytoriów (zarządzanie kontami, linkami DOI, migracje danych).





Cel archiwizacji – korzyści dla przyszłych badań i rozwoju nau

Czy tylko spełnienie wymagań?

Przypadki ponownego użycia (poza kontekstem pierwotnych badań)

- weryfikacja przez innych

Archiwizacja danych umożliwia innym badaczom weryfikację wyników badań. Dzięki dostępowi do oryginalnych danych, mogą oni przeprowadzić niezależne analizy, co zwiększa wiarygodność i rzetelność badań

- dodatkowe analizy w przyszłości

Zarchiwizowane dane mogą być ponownie wykorzystane do dalszej analizy. Nowe technologie lub metody analizy mogą pozwolić na odkrycie nowych wniosków z już istniejących danych

- budowanie reputacji akademickiej

Udostępnienie zarchiwizowanych danych może przyczynić się do budowania reputacji akademickiej badaczy. Publikowanie danych w otwartych repozytoriach zwiększa widoczność i cytowalność ich pracy

- kolejne publikacje

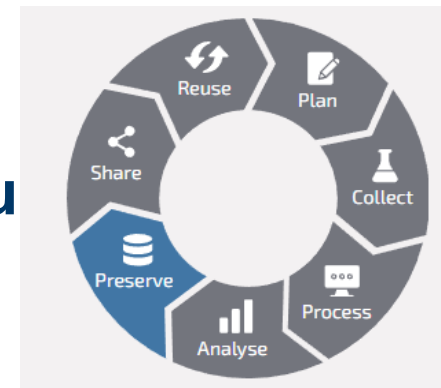
Zarchiwizowane dane mogą być podstawą do dalszych publikacji naukowych. Badacze mogą wykorzystać istniejące dane do tworzenia nowych artykułów, książek, raportów itp..

- nauczanie i uczenie się

Zarchiwizowane dane mogą być wykorzystywane w procesie nauczania i uczenia się. Studenci i nauczyciele mogą korzystać z rzeczywistych danych badawczych do nauki metod analizy, statystyki czy innych aspektów badań naukowych

- ponowne użycie (także prywatne)

Badacze mogą również korzystać z archiwizowanych danych do przyszłego prywatnego użytku. Mogą wrócić do swoich danych w celu dalszej analizy lub wykorzystania ich w nowych projektach badawczych



Dobre praktyki przygotowania danych do archiwizacji długoterminowej

Dokumentacja danych :

- o *opis struktury plików i zbiorów danych*
- o *opis zmiennych, jednostek pomiarowych, kodów użytych w danych*
- o *notatki i kontekst badań*

Standaryzacja formatów plików:

- o *Stosowanie formatów otwartych i trwałych (np. CSV zamiast XLSX, TIFF zamiast JPEG, PDF/A zamiast DOCX)*

Organizacja danych i struktura katalogów:

- o *jasna i logiczna struktura folderów (np. dane_surowe, dane_przetworzone, kod, dokumentacja)*
- o *spójne nazewnictwo plików (np. wg daty, wersji, typu danych)*

Zarządzanie wersjami danych:

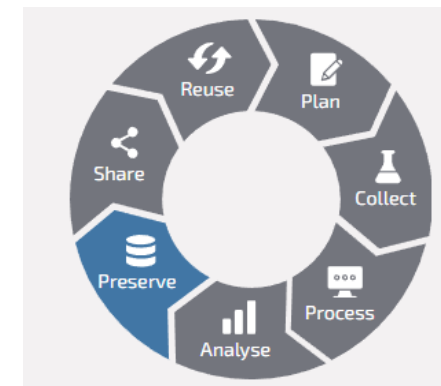
- o *Dokumentowanie zmian w danych (np. plik README lub changelog)*
- o *Wersjonowanie ręczne lub za pomocą narzędzi*

Metadane:

- o *Opis danych zgodnie z uznanymi standardami*
- o *Dodanie informacji o licencjach, autorach, dacie utworzenia*

Kontrola jakości i integralności danych:

- o *Sprawdzenie spójności, kompletności, braku duplikatów*
- o *Weryfikacja zgodności danych z dokumentacją*





Lista sprawdzająca

Pliki/zestawy danych :

- Czy wszystkie dane końcowe zostały zidentyfikowane do archiwizacji?
- Czy pliki zostały zapisane w trwałych, otwartych formatach (np. CSV, TIFF, PDF/A)?
- Czy pliki/foldery zostały nazwane zgodnie z określonym standardem nazewnictwa plików?

Dokumentacja i metadane:

- Czy dane zawierają plik README z informacjami kontekstowymi (kto, co, kiedy, jak)?
- Czy zastosowano standardowe schematy metadanych ?
- Czy dokumentacja opisuje jednostki miar, kody kategorii, brakujące dane itp.?

Lokalizacja archiwizacji i udostępnienie:

- Czy wybrano certyfikowane repozytorium danych (np. Most Danych)
- Czy dostęp do danych został określony (otwarty / ograniczony / embargo)?
- Czy nadano trwały identyfikator (np. DOI)?

Zgodność z przepisami

- Czy dane nie naruszają praw autorskich lub umów licencyjnych?
- Czy dane osobowe są odpowiednio zanonimizowane lub pseudonimizowane?
- Czy zapewniono zgodność z RODO i krajowymi przepisami?

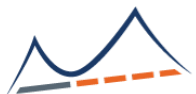
Bezpieczeństwo i przechowywanie:

- Czy zaplanowano długoterminowe utrzymanie danych?
- Czy dane posiadają kopię bezpieczeństwa?
- Czy zabezpieczono budżet na archiwizację?
- Czy zabezpieczono budżet na archiwizację?



POLITECHNIKA
GDAŃSKA

ARCHIWIZACJA, PRZECHOWYWANIE, KOPIA ZAPASOWA - narzędzia na PG



MOST DANYCH

Repozytorium z certyfikatem CoreTrustSeal

- Archiwizacja
- Powszechne Udostępnianie
- wersjonowanie
- Długoterminowe przechowywanie
- Darmowa przestrzeń
- Kopie zapasowe udostępnionych danych
- Metadane



Office 365 (OneDrive, Teams, SharePoint)

- Przechowanie podczas projektu
- Synchronizacja danych
- Współdzielenie/współpraca
- Kopia zapasowa (ręczna lub automatyczna)
- Przestrzeń dyskowa zależna od licencji (koszt)

Więcej: <https://cui.pg.edu.pl/wsparcie-techniczne>

ci task

Sieć współpracy
Moc obliczeń

Przechowywanie i przetwarzanie dużych zbiorów danych

- Archiwizacja
- Ograniczone udostępnianie
- Automatyczne kopie zapasowe
- Długoterminowe przechowywanie
- Więcej: <https://task.gda.pl/pl/uslugi/dla-naukowcow/>

Usługa CI TASK - NCNdata

w okresie realizacji projektu:

- 100 GB przestrzeni składowania danych,
- dostęp do danych dla 5 członków zespołu badawczego za pomocą indywidualnych kont,
- dostęp do danych za pomocą aplikacji na urządzenia końcowe oraz przez przeglądarkę internetową,
- automatyczny mechanizm tworzenia kopii zapasowych danych: wykonywane raz w tygodniu z dwutygodniową retencją.
- po zakończeniu projektu przechowywana jest jedna kopia zapasowa.
- Dodatkowa przestrzeń 40 zł / TB / miesiąc (1 TB przestrzeni obejmuje 342 GB przestrzeni przechowywania oraz 2 x 342 GB przestrzeni na dwie kopie zapasowe)

Zasady korzystania:

https://task.gda.pl/files/zasady_swadczenia_i_korzystania_z_uslugi_NCNdata.pdf



- [Centrum Kompetencji Otwartej Nauki | Politechnika Gdańska](#)
 - [Zarządzanie danymi badawczymi | Politechnika Gdańska](#)
- [Research Data Management Toolkit](#)
- [MANTRA Research Data Management Training](#)
- [CI TASK – dla naukowców](#)
- [CI TASK – NCNData](#)
- [Dane Badawcze - Katalog Danych Badawczych - MOST Wiedzy](#)

Szkolenie finansowane w ramach "Działań na rzecz promowania i zwiększenia świadomości w zakresie Otwartych Danych Badawczych", które stanowią element realizacji programu „Inicjatywa Doskonałości – Uczelnia Badawcza” (IDUB) w zakresie podniesienia poziomu jakości działalności naukowej uczelni w ramach Działania I.4 (Działania na rzecz zwiększenia liczby publikacji w prestiżowych czasopismach i wydawnictwach).