Repozytoria danych badawczych

Wojciech Fenrich

Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego
Uniwersytet Warszawski

wf@icm.edu.pl

Zadanie realizowane przez Narodowe Centrum Nauki na podstawie zlecenia Ministra Edukacji i Nauki dot. krajowej koordynacji partnerstwa European Open Science Cloud w latach 2022-2023.

Co to jest repozytorium?

- Horizon Europe AGA: Repozytorium to internetowe archiwum, w którym badacze mogą deponować cyfrowe rezultaty działalności badawczej i zapewnić do nich (otwarty) dostęp.
- Repozytoria mogą gromadzić:
  - Publikacje
  - Dane badawcze - i na tych się dziś skupimy…

Dane badawcze, ale w jakiej postaci?

- Repozytoria gromadzą zbiory danych (ang. *data sets*), na które składają się:
  - metadane (czyli dane opisujące dane),
  - dokumentacja,
  - pliki z danymi.

Typy repozytoriów danych badawczych

- Repozytoria specjalistyczne, np. PDB.
- Repozytoria dziedzinowe, np. GESIS, Pangaea.
- Repozytoria instytucjonalne, np. Cambridge, RODBUK.
- Repozytoria ogólnego przeznaczenia (sieroce), np. Zenodo, RepOD.

Zakres weryfikacji danych
(Za: CoreTrustSeal AMT)

1. **Brak weryfikacji** – treści są publikowane w takiej postaci, w przychodzą do repozytorium.
2. **Poziom podstawowy** – zgrubne sprawdzenie poprawności elementów zbioru, dodanie podstawowych metadanych lub dokumentacji.
3. **Poziom rozszerzony** – konwersja do nowych formatów, rozszerzenie dokumentacji.
4. **Poziom danych** – dodatkowo dokonywane jest sprawdzenie i edycja danych.

Kwestia zobowiązań

- Warto zwrócić uwagę na to, czy nasz grantodawca, pracodawca lub czasopismo nie posiada określonych preferencji względem miejsca zdeponowania danych.
- Tych należy szukać w odpowiednich politykach zarządzania danymi.
- Preferencje takie mogą dotyczyć konkretnych serwisów, ale też serwisów jakiegoś konkretnego rodzaju (np. repozytorium dziedzinowe czy zaufane repozytorium).

Kwestia zobowiązań – uwaga na marginesie

- Tworząc politykę zarządzania danymi warto przewidzieć w niej przypadek zbiegu zobowiązań.

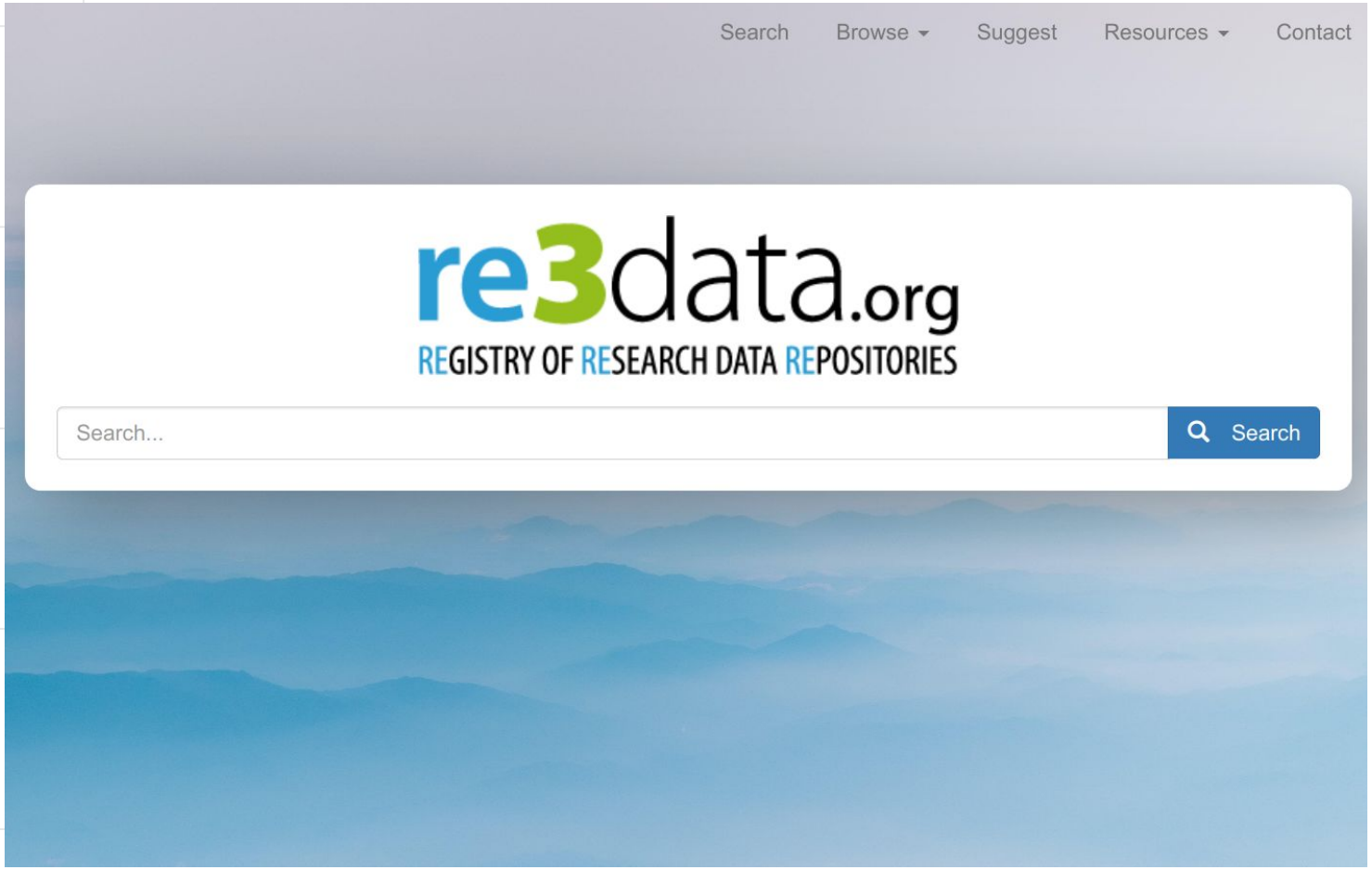
- Zbiegające się zobowiązania mogą ze sobą kolidować, a polityka powinna określać, co należy zrobić w takiej sytuacji.

Jaki typ repozytorium wybrać?

- Zacznijmy poszukiwania na jak najniższym poziomie, najlepiej od repozytorium specjalistycznego.
- W razie niepowodzenia – poszukajmy “oczko” wyżej.

Gdzie szukać odpowiedniego repozytorium?

https://re3data.org/



Zaufane repozytoria – kryteria NCN

1. Repozytoria posiadające certyfikat jakości (CoreTrustSeal, DIN-31644, ISO-16363 lub WDS/ICSU).
2. Repozytoria szeroko uznane w danej dyscyplinie.
3. Kryteria zawarte w dokumencie *Practical Guide to the international alignment of research data management*
    a. Trwałe identyfikatory.
    b. Metadane.
    c. Dostęp do danych oraz licencje.
    d. Długotrwałe przechowywanie.

Zaufane repozytoria – Horizon Europe

- Repozytoria certyfikowane (CoreTrustSeal, DIN31644, ISO16363).
- Repozytoria dziedzinowe powszechnie wykorzystywane i wspierane przez społeczność badaczy.
- Repozytoria instytucjonalne i ogólnego przeznaczenia:
  - charakteryzujące się odpowiednim poziomem technicznym i organizacyjnym,
  - zapewniające darmowy, szeroki, najlepiej otwarty dostęp do treści,
  - zapewniające trwałe identyfikatory, maszynową dostępność metadanych w odpowiednich standardach,
  - wspierają średnio– i długoterminowe przechowywanie zdeponowanych danych.

Im wcześniej, tym lepiej

- Trafny wybór repozytorium na wczesnym etapie projektu powoduje, że zarządzanie danymi staje się dużo łatwiejsze.
- Odpowiednie repozytorium to miejsce, w którym wiele osób dysponujących danymi zbliżonymi do naszych dawno rozwiązało problemy, przed którymi stoimy lub za moment staniemy.

Co nam daje dobrze dobrane repozytorium?

- Odpowiedni schemat metadanych.
- Trwały identyfikator (najczęściej DOI).
- Szereg materiałów dodatkowych (materiały informacyjne, szkoleniowe, narzędzia etc.).
- Wsparcie ze strony kompetentnych osób.

Po co komu trwałe identyfikatory?

- Dużo łatwiejsze gromadzenie informacji o cytowaniach: wiele stylów cytowania – identyfikator zawsze taki sam.
- Trwałość dostępu do danych: cytując z wykorzystaniem trwałego identyfikatora uodparniamy cytowanie na zmiany w jego adresie URL (przenosiny, zmiana oprogramowania).

Trzy stany numerów DOI

- **Draft** – może zostać usunięty, charakterystyczny dla wersji roboczych zbiorów danych. Stan może zostać zmieniony na “Findable” lub “Registered”, ale nie ma do niego powrotu. Nie jest to numer DOI w ścisłym sensie, bo jego “rozwiązanie” nie jest możliwe.
- **Findable** – możliwy do “rozwiązania”, jest indeksowany w publicznej infrastrukturze Datacite. Może zostać zmieniony na “Registered”.
- **Registered** – możliwy do “rozwiązania”, nie jest indeksowany w publicznej infrastrukturze Datacite, charakterystyczny dla wycofanych zbiorów danych. Może zostać zmieniony na “Findable”.

Na czym polega trwałość trwałego identyfikatora

- Trwały identyfikator zawsze powinien wskazywać na jakiś obiekt.
- W przypadku DOI dla zbiorów danych: “landing page” działającej strony internetowej zawierającej informacje o zbiorze danych.
- Jeśli zaistnieje potrzeba wycofania jakiegoś zbioru danych (tak by przestał być publicznie dostępny), DOI nadal powinno wskazywać na stronę z tzw. “nagrobkiem” (ang. tombstone).

DOI – rzut oka za kulisy

Required Properties

*** DOI** The globally unique string that identifies the resource and can't be changed.

10.18150/8527338

*** State** The state determines whether a DOI is registered and findable. Once in Registered or Findable state, a DOI can't be set back to Draft state. More …

- Draft only visible in Fabrica, DOI can be deleted
- Registered registered with the DOI Resolver
- Findable registered with the DOI Resolver and indexed in DataCite Search

*** URL** The location of the landing page with more information about the resource.

https://repod.icm.edu.pl/dataset.xhtml?persistentId=doi:10.18150/8527338

Should be a https URL — within the allowed domain(s) of your repository if domain restrictions are enabled in the repository settings. Http and ftp are also supported. For example http://example.org

Polskie repozytoria danych badawczych

- CLARIN – https://clarin-pl.eu/dspace/
- Open Forest Data – https://openforestdata.pl/
- Most Danych – https://mostwiedzy.pl/pl/
- Repozytorium Danych Społecznych – https://rds.icm.edu.pl/
- MX-RDR – https://mxrdr.icm.edu.pl/
- RepOD – https://repod.icm.edu.pl/
- RODBUK – https://rodbuk.pl/

Zagraniczne repozytoria danych badawczych

Zenodo – https://zenodo.org/

Harvard Dataverse – https://dataverse.harvard.edu/

Figshare – https://figshare.com/

Dryad – https://datadryad.org/

Dodatkowe funkcje (niektórych) repozytoriów

- Embargo.
- Prywatny link.
- Możliwość udostępniania danych w ograniczonym dostępie.
- Możliwość zarejestrowania metadanych dla zbioru danych, którego udostępnienie nie jest planowane.

A co z kodem?

- Naturalnym miejscem dla kodu jest serwis GitHub (https://github.com/).
- Po numer DOI możemy udać się do repozytorium Zenodo.
- Tym, co otrzymuje numer DOI jest pojedynczy *release.*
- Nowy *release* – nowy numer DOI.
- Szczegóły: https://docs.github.com/en/repositories/archiving-a-github-repository/referencing-and-citing-content

Gdzie szukać danych?

- OpenAIRE Explore: https://explore.openaire.eu/
- Google Dataset Search: https://datasetsearch.research.google.com/
- Harvard Dataverse: https://dataverse.harvard.edu/

Data journals jako uzupełnienie repozytoriów

- Publikują artykuły poświęcone danym naukowym. Same dane są deponowane w odpowiednich repozytoriach.
- Włączenie informacji o danych w ramach ekosystemu publikacji naukowych.
- Rozwiązanie problemu “spóźnionego ewaluatora”.

Często zadawane pytania: *Czy to repozytorium zapewnia zgodność z zasadami FAIR?*

- Repozytorium zapewnia zgodność z zasadami FAIR w pewnych aspektach, przede wszystkim technicznych.
- Nawet w repozytorium zgodnym z zasadami FAIR można udostępnić zbiór niezgodnie z zasadami FAIR.
  - Np. repozytorium *umożliwia wprowadzenie* odpowiednich metadanych dziedzinowych, ale osoba deponująca dane ich *nie wprowadza*.
- Niektóre repozytoria mogą jednak oferować znacznie szerszy zakres usług, a wtedy zapewnienie zgodności z zasadami FAIR może wykraczać poza aspekt techniczny.

Często zadawane pytania: *Czy możemy prosić o priorytetowe zajęcie się naszym zbiorem?*

- Tego typu pytania padają często w okresach raportowania.
- Wtedy jednak zadaje je wiele osób naraz.
- Dane warto zdeponować z odpowiednim wyprzedzeniem.

Często zadawane pytania: *Czy mogę prosić o wycofanie mojego zbioru danych?*

- Na zbiór danych warto spojrzeć podobnie jak na artykuł naukowy. Nie jest to to samo, co “wrzucenie plików do chmury”.
- Wycofanie zbioru danych powinno być zdarzeniem wyjątkowym.
- Słabo uzasadnione pytanie o wycofanie zbioru danych może świadczyć o tym, że decyzja o zdeponowaniu danych była pochopna.

Jak wygląda deponowanie danych w repozytorium na przykładzie repozytorium RepOD

[Film]

Dziękuję za uwagę

Wojciech Fenrich, ICM UW
wf@icm.edu.pl

GRAMY DLA POLSKIEJ NAUKI

Zadanie realizowane przez Narodowe Centrum Nauki na podstawie zlecenia Ministra Edukacji i Nauki dot. krajowej koordynacji partnerstwa European Open Science Cloud w latach 2022-2023.