Zarządzanie danymi badawczymi w naukach medycznych, naukach farmaceutycznych i naukach o zdrowiu

Szymon Kubik, Uniwersytet Jagielloński Collegium Medicum

Jakub Rusakow, Gdański Uniwersytet Medyczny

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.

Dlaczego zarządzanie danymi badawczymi jest ważne w naukach medycznych, farmaceutycznych i o zdrowiu

„DATA IS THE NEW OIL”

Clive Humby, 2006

Charakterystyka danych w naukach medycznych, farmaceutycznych i naukach o zdrowiu

1. **Pochodzące z badań empirycznych** - przedklinicznych, klinicznych, eksperymentów medycznych, genetycznych…
2. **Szereg ograniczeń i wymogów**: prawnych, etycznych, metodologicznych, epidemiologicznych, organizacyjnych…
3. **Duże ilości danych** – konieczność redukcji, grupowania i selekcji danych;
4. **Dane surowe** -> **dane przetworzone;**
5. **Różnorodność** rodzajów danych, formatów i rozmiarów plików;

Charakterystyka danych w naukach medycznych, farmaceutycznych i naukach o zdrowiu – c. d.

6. **Bezpośredni wpływ** na społeczeństwo;
7. **Często bardzo „medialne”** – wyniki badań cytowane nie tylko przez środowisko naukowe, ale też media tradycyjne i nowe media;
8. **Szybko starzejące się** – rozwój nauki sprawia, że dane dość szybko tracą na swej aktualności;
9. **Generowane przy dużym nakładzie finansowym** – aparatura badawcza, laboratoria, banki danych.

Zasady FAIR

F

Findable: odnajdywalne

- Dane powinny być dokładnie i szczegółowo opisane poprzez **metadane** – tak, by możliwe było ich odnalezienie za pomocą wyszukiwarek;
- Danym powinien zostać przyporządkowany **unikalny identyfikator** – umożliwiający jednoznaczne wskazanie na zbiór danych;
- Dane powinny zostać zdeponowane i udostępnione za pośrednictwem narzędzi umożliwiających ich odnalezienie – **repozytoria danych badawczych, *data journals***.

Zasady FAIR

A

Accessible: dostępne

- Dane powinny pozostać dostępne w wybranym repozytorium lub czasopiśmie i za pośrednictwem identyfikatora;
- Dostęp do danych nie powinien wymagać dodatkowych narzędzi ani oprogramowania;
- Jeśli dane przestają być dostępne, powinien zostać zachowany **dostęp do metadanych jasno opisujących dane**.

Zasady FAIR

Interoperable: interoperacyjne

- Metadane opisujące dane w odpowiednim **standardzie** oraz z wykorzystaniem **słownictwa formalnego** stosowanego w naukach medycznych – np. MeSH;
- **Możliwe do odczytania** zarówno przez ludzi, jak i przez maszyny;
- Możliwość **powiązania** ich z innymi zasobami – innymi danymi, publikacjami…

Zasady FAIR

R

Reusable: możliwe do ponownego wykorzystania

- Im lepiej **opisane dane** (wiele dokładnych atrybutów metadanych), tym większa szansa na ich ponowne wykorzystanie;
- Odpowiednio przygotowana **dokumentacja** z informacjami o autorze danych, pochodzeniu danych i ich przetwarzaniu;
- Określenie **licencji** zezwalającej na korzystanie z danych.

Zasady otwierania danych badawczych

Dane powinny być tak otwarte, jak to możliwe i na tyle zamknięte, na ile to jest konieczne.

[jęz. ang.: *as open as possible, as closed as necessary*]

Uzasadnione przypadki ograniczania dostępu do danych badawczych:

1. Dane badawcze zawierają **dane osobowe** i umożliwiają dostęp do **danych wrażliwych** (np. pacjentów), zaś ich anonimizacja czy pseudonimizacja znacząco wpływa na ich kształt;
2. Uczestnicy badania **nie udzielili zgody** na udostępnianie danych;
3. Udostępnienie danych badawczych budzi poważne **wątpliwości natury etycznej**;
4. Dane objęte są **ochroną patentową**, **tajemnicą handlową** lub **ochroną informacji niejawnych**.

Jeżeli organizacja finansująca badanie stosuje politykę otwartego dostępu to **dane powiązane z publikacją** powinny zostać otwarte.

Plan Zarządzania Danymi
Ogólne Informacje

Dlaczego Plan Zarządzania Danymi jest ważny w naukach medycznych, farmaceutycznych i o zdrowiu?

- Wcześniej zaplanowane wykorzystanie istniejących już danych może znacznie ograniczyć koszty projektu;
- Plan pozwala oszacować koszty zarządzania danymi (np. koszty przechowywania, które mogą być spore);
- Zabezpieczenie nieraz bardzo kosztownych danych;
- Kontrola jakości danych;
- Wyjątkowo dużo danych wrażliwych (dane osobowe pacjentów, uczestników badania);
- Zgodność z regulacjami prawnymi, regulaminami – uniknięcie odpowiedzialności;
- Wcześniejszy wybór repozytorium pozwala nam zaplanować odpowiednie standardy metadanych;
- Zaplanowanie sposobu dostępu do danych może ułatwić współpracę pomiędzy naukowcami pracującymi w danym projekcie;
- Zaplanowanie publikacji zgodnie z wymogami wydawców.

Opis danych oraz pozyskiwanie lub ponowne wykorzystanie dostępnych danych

Opis danych

Sposób pozyskiwania lub wytwarzania nowych danych lub ponownego wykorzystywania danych istniejących.

Dane pierwotne (nowe):

- W jaki sposób zbieramy, pozyskujemy nowe dane?
- Jakiego rodzaju to są dane (w skrócie)?
- Za pomocą jakiego sprzętu i oprogramowania? Jakie standardy są stosowane?
- Jak będziemy zapisywać i porządkować te dane (w skrócie)?

Dane wtórne (istniejące wcześniej):

- Do kogo należą te dane?
- Skąd pochodzą? Czy ich wartość i jakość jest wystarczająco udokumentowana?
- W jaki sposób zostały zebrane, wytworzone?
- Na jakiej zasadzie będzie się odbywać pozyskiwanie tych danych (licencje)?
- Czy konieczne są dodatkowe prace na tych danych (oczyszczanie, segregacja, digitalizacja)?

Opis danych

Jakie dane będą pozyskiwane lub wytwarzane w projekcie (rodzaj, format, ilość)?

Przykładowe rodzaje danych:

- Zdjęcia (np. RTG, TK, USG);
- Dane statystyczne, dane tekstowe;
- Ankiety;
- Dane pomiarowe;
- Próbki fizyczne.

Formaty danych (otwarte vs. zamknięte, .csv vs. .xlsx) i związane z tym oprogramowanie (zamknięte lub otwarte) :

- Formaty właściwe dla sprzętu i oprogramowania;
- Szacunkowa objętość danych wytworzonych, pozyskanych, ponownie wykorzystanych (ile MB, GB, TB)?

Dokumentacja i jakość danych

Dokumentacja i jakość

Metadane i dokumentacja dot. danych w naukach medycznych, naukach o zdrowiu i naukach farmaceutycznych

- **METADANE** – ich zakres i standardy (np. Darwin Core);
- Informacje, których sami byśmy potrzebowali;
- Zgodność ze standardem repozytorium (repozytoria specjalistyczne);
- Maszynowy odczyt metadanych (machine-readable);
- **README, ORCID;**
- Dokumentacja: szeroki opis metodologii (techniki, opis próby, labbooki, zgody uczestników);
- Wersjonowanie datasetów, sposób organizacji plików i folderów;

Dokumentacja i jakość

Środki kontroli jakości danych

- Infrastruktura badawcza – certyfikaty, atesty, sposoby kalibracji;
- Laboratoria – infrastruktura, certyfikaty;
- Kwalifikacje osób wykonujących pomiary, wprowadzających dane, dokonujących obliczeń;
- Stosowane w dyscyplinie standardy (np. ICH, WHO);
- Kwestionariusze ankiet – autorskie i standaryzowane;
- SOP, regulaminy, wewnętrzne wytyczne;
- Kontrola dostępu do danych (zabezpieczenie przed nieuprawnioną modyfikacją);
- Eliminacja błędów pomiarowych i stronniczości (bias), np. za pomocą standardowego protokołu dla zespołu, walidacji przez różnych członków zespołu.

Przechowywanie i tworzenie kopii zapasowych podczas badań

Przechowywanie i kopie zapasowe

Przechowywanie danych i metadanych

- Gdzie i w jaki sposób będą przechowywane dane podczas projektu?
- Potencjalnie niebezpieczne sytuacje mogą ce skutkować utratą danych;
- Procedury tworzenia kopii zapasowych, powiązana z infrastrukturą, np. zasada 3-2-1;
- Odzyskiwanie danych w przypadku utraty/uszkodzenia;
- Bezpieczeństwo przepływu danych między członkami zespołu;
- Przenoszenie danych, digitalizacja w celu zabezpieczenia i rozpowszechnienia;
- Częstotliwość wykonywania kopii, odpowiedzialność.

Przechowywanie i kopie zapasowe

W jaki sposób zostanie zapewnione bezpieczeństwo i ochrona danych wrażliwych w okresie trwania projektu?

- Bezpieczeństwo każdego rodzaju danych (nie tylko wrażliwych)
- Sposób i miejsce przechowywania danych wrażliwych/osobowych – szyfrowanie, zabezpieczenia IT, hasła;
- Sposób odzyskiwania danych utraconych w wyniku incydentu (kopia zapasowa);
- Kontrola dostępu do danych – wypracowanie systemu (zwłaszcza w przypadku współpracy między kilkoma partnerami z kraju lub zagranicy).

Wymogi prawne, kodeksy postępowania

Wymogi prawne, kodeksy postępowania

Jeżeli będzie miało miejsce przetwarzanie danych osobowych, w jaki sposób zostanie zapewniona zgodność z przepisami dotyczącymi danych osobowych oraz ich ochrony? [jeżeli dotyczy]

- Czy to już dane osobowe?
- Kiedy anonimizacja lub pseudonimizacja danych jest niezbędna – przykłady;
- Regulacje prawne UE – RODO (GDPR), krajowe – UODO, instytucjonalne – zarządzenie Rektora, Uchwała Senatu, kodeksy etyki;
- Obligatoryjna zgoda uczestnika na udział w badaniu (w jakiej formie i jak zabezpieczona).

W przypadku pełnej anonimizacji, zgoda na udostępnianie danych nie jest konieczna, ale jest to elementem kultury badawczej i dobrej praktyki.

Zgoda powinna uwzględniać:

- informację o archiwizowaniu danych badawczych i osobowych (dane osobowe będą bezpiecznie przechowywane/będą niejawne i ostatecznie zniszczone a zanonimizowane dane badawcze będą jawne, udostępniane innym i przechowywane przez czas określony w wytycznych organizacji finansującej badanie);
- informację o długoterminowym przechowywaniu i wykorzystaniu zanonimizowanych danych;
- oświadczenie uczestnika badania potwierdzające świadomą zgodę na udział.

Wymogi prawne, kodeksy postępowania

W jaki sposób zostanie zapewniona zgodność z innymi przepisami, takimi jak prawa własności intelektualnej i prawa własności? Jakie przepisy mogą znaleźć zastosowanie w przypadku omawianej dyscyplin/dziedziny?

- Określenie, do kogo należą już wytworzone dane i do kogo będą należeć dane wytworzone w projekcie;
- Twórca a właściciel danych;
- Zastosowanie licencji (Creative Commons);
- Ograniczenia ponownego wykorzystania danych pochodzących od osób trzecich;
- Wskazanie regulacji prawnych – zewnętrznych i wewnętrznych;

Wymogi prawne, kodeksy postępowania

- Prawo autorskie – na ogół nie ma zastosowania do danych badawczych
- Prawo własności intelektualnej
- Zarządzenia Rektora/Uchwały Senatu
- Polityka Otwartości Instytucji
- Ustawa o ochronie baz danych
- Umowy dwustronne i konsorcyjne
- Regulamin konkursu
- Umowa o grant
- Licencje (np. oprogramowanie / sprzęt)
- Regulamin korzystania z repozytorium

Udostępnianie i długotrwałe przechowywanie danych

Udostępnianie i długotrwałe przechowywanie

Kiedy i w jaki sposób będą udostępniane dane z projektu? Czy istnieją ewentualne ograniczenia i zakazy dotyczące ich udostępniania?

1. **Określenie sposobu przekazania informacji o danych** potencjalnym użytkownikom:
   - repozytorium danych,
   - czasopismo danych,
   - informacja w publikacji;
2. **Wskazanie czasu udostępnienia danych**:
   - w trakcie trwania projektu – wersjonowanie,
   - przed publikacją pracy naukowej,
   - w momencie publikacji pracy naukowej,
   - po publikacji pracy naukowej – nałożenie embargo na dane (np. 3 miesiące, rok, 2 lata…);

Udostępnianie i długotrwałe przechowywanie

Kiedy i w jaki sposób będą udostępniane dane z projektu? Czy istnieją ewentualne ograniczenia i zakazy dotyczące ich udostępniania?

3. **Ustalenie sposobu i okresu przechowywania danych**:
   - w trakcie trwania projektu,
   - po zakończeniu projektu – co najmniej 10 lat,
   - metadane dostępne bezterminowo;
4. **Wskazanie ograniczeń i przeszkód w możliwości pełnego lub częściowego udostępnienia danych** – konieczne uzasadnienie:
   - prawne,
   - etyczne,
   - infrastrukturalne,
   - finansowe;

Udostępnianie i długotrwałe przechowywanie

Kiedy i w jaki sposób będą udostępniane dane z projektu? Czy istnieją ewentualne ograniczenia i zakazy dotyczące ich udostępniania?

5. **Zaplanowanie drogi udostępnienia wyników badań**:
   - Czy wydawcy czasopism wymagają udostępniania danych powiązanych z publikacją?
   - Czy wydawcy wskazują gdzie należy udostępnić dane?
   - Czy repozytoria brane pod uwagę są certyfikowane?
   - Czy zdeponowanie w nich danych będzie wymagało dodatkowych opłat?
6. **Informacja czy udostępnianie danych wymaga zgody uczestników badania.**

Udostępnianie i długotrwałe przechowywanie

Jak będzie wyglądać selekcja danych przeznaczonych do utrwalenia i gdzie będą one długoterminowo przechowywane?

1. **Selekcja danych przeznaczonych do utrwalenia (dot. danych istotnych)**
   - dane niedostępne publicznie,
   - dane pozyskane za świadomą zgodą uczestnika badań,
   - dane pozbawione danych personalnych lub wrażliwych
     - jeśli możliwe - anonimizacja,
   - dane pozyskane za zgodą komisji etycznych;
2. **Wskazanie danych, które należy zachować oraz tych, które należy zniszczyć**
   - umowy,
   - przepisy prawne,
   - regulacje prawne

Udostępnianie i długotrwałe przechowywanie

Jak będzie wyglądać selekcja danych przeznaczonych do utrwalenia i gdzie będą one długoterminowo przechowywane?

1. Organizacja danych zgodna z wytycznymi FAIR;
2. Wskazanie metody nazewnictwa plików i ich wersjonowania;
3. Wybór miejsca zdeponowania i udostępniania danych (wybór repozytorium)

Kryteria wyboru repozytorium

Typy repozytoriów:

- **Repozytoria ogólne:**
  - Zenodo (https://zenodo.org/),
  - RepOD (https://repod.icm.edu.pl),
  - Mendeley Data (https://data.mendeley.com/)
- **Repozytoria dziedzinowe:**
  - European Nucleotide Archive (https://www.ebi.ac.uk/ena/browser/home),
  - Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/),
  - Polska Platforma Medyczna (https://ppm.edu.pl/)

Kryteria wyboru repozytorium

- **Repozytoria o wysokiej reputacji dla dziedziny/dyscypliny**
  - popularność repozytorium w społeczności naukowców z naszej dyscypliny,
  - repozytoria, w których sami poszukujemy danych;

- **Repozytoria posiadające certyfikat CoreTrust Seal;**

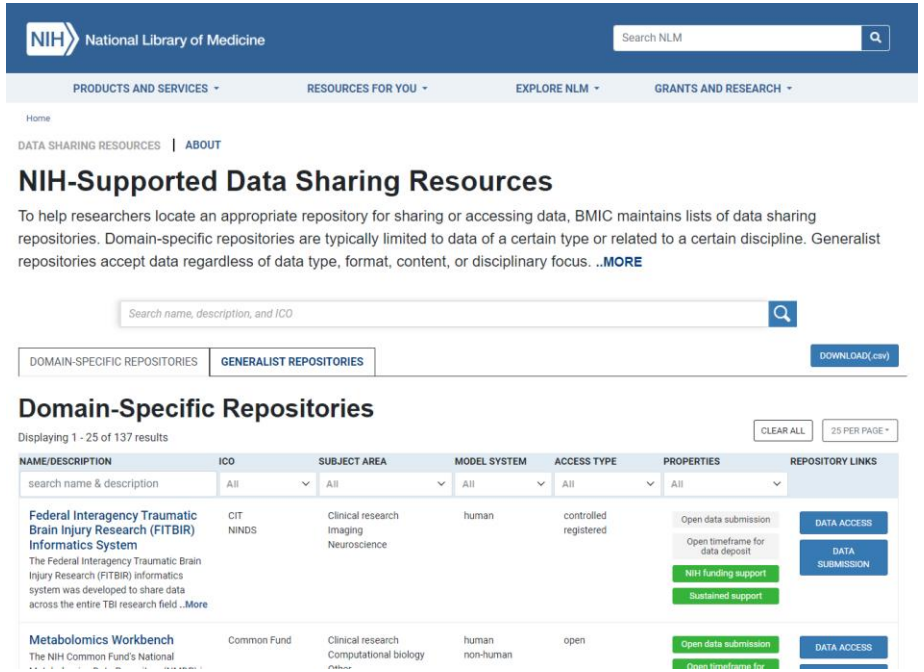
- **Repozytoria z zasobami rejestrowanymi w bazach indeksujących dane badawcze:**
  - Web of Science Data Citation Index,
  - Elsevier Data Monitor,
  - Mendeley Data,
  - Google Dataset Search

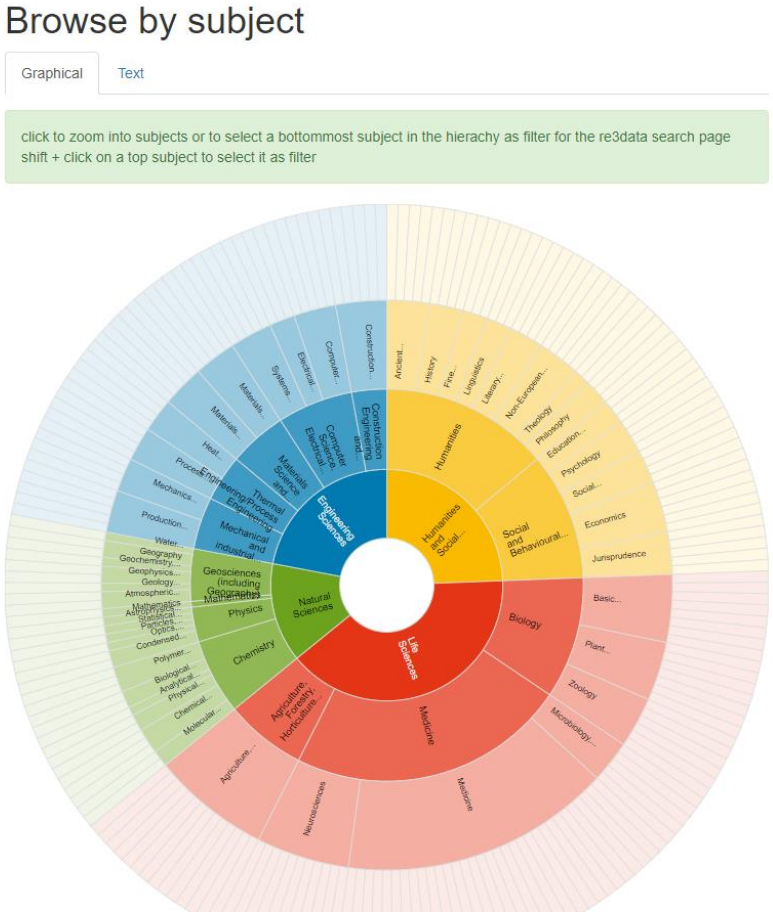
Kryteria wyboru repozytorium

- **Repozytoria spełniające warunki formalne:**
  - stosowanie zasad FAIR,
  - stosowanie unikalnego identyfikatora zasobu (PID):
    - DOI,
    - Handle...,
  - stosowanie otwartych licencji,
  - czas przechowywania danych:
    - Dane co najmniej 10 lat,
    - metadane bezterminowo,
  - możliwość zastosowania embargo,
  - techniczne możliwości repozytorium:
    - formaty plików,
    - rozmiary plików

Kryteria wyboru repozytorium

- **Rejestry repozytoriów:**
  - Registry of Research Data Repositories (https://www.re3data.org/)
  - NIH Data Sharing Repositories (https://www.nlm.nih.gov/NIHbmic/)





Udostępnianie i długotrwałe przechowywanie

Jakie metody lub oprogramowanie umożliwiają dostęp do danych i korzystanie z danych?

- **Mechanizm udostępniania danych:**
  - dostęp poprzez repozytorium,
  - odpowiedzi na żądanie,
- **Przekształcenie danych do formatu standardowego lub otwartego:**
  - aby mogły być przechowywane w repozytorium danych badawczych,
  - Aby mogły być przechowywane w sposób długotrwały,
  - Aby zachowały długi okres ważności
- **informacja czy do skanowania lub konwersji niezbędne są dodatkowy sprzęt lub oprogramowanie**

Udostępnianie i długotrwałe przechowywanie

W jaki sposób zagwarantować stosowanie unikalnego i trwale przypisanego identyfikatora [ang. PID] dla każdego zbioru danych?

- **Wybór właściwego repozytorium danych badawczych - nadającego trwały identyfikator**

Zadania związane z zarządzaniem danymi oraz zasoby

Zadania związane z zarządzaniem danymi oraz zasoby

Podział ról w projekcie badawczym:

- kto wytwarza dane badawcze,
- kto odpowiada za jakość danych,
- kto odpowiada za archiwizację i długoterminowe zarządzanie danymi,
- kto tworzy i wdraża oraz wersjonuje Plan Zarządzania Danymi,

Zadania związane z zarządzaniem danymi oraz zasoby

Wsparcie dla procesu zarządzania danymi badawczymi w jednostce badawczej:

- data stewardzi,
- działy projektów,
- bibliotekarze,
- zespół IT,
- rzecznicy patentowi,
- radcy prawni,
- ...

Zadania związane z zarządzaniem danymi oraz zasoby

Budżet na cele zarządzania danymi i zagwarantowanie przestrzegania zasad FAIR

- dodatkowe zasoby do zarządzania danymi:
  - osoby,
  - czas,
  - sprzęt,
  - oprogramowanie;
- koszty związane z zapewnieniem standardów FAIR w projekcie

Zadania związane z zarządzaniem danymi oraz zasoby

Budżet na cele zarządzania danymi i zagwarantowanie przestrzegania zasad FAIR

- Koszty pośrednie OA – w wysokości do 2% przyznanych kosztów bezpośrednich

  Do wykorzystania **wyłącznie** na koszty związane z udostępnieniem publikacji i/lub danych badawczych w otwartym dostępie

- Pozostałe koszty pośrednie - w wysokości do 20% od wykorzystanych kosztów bezpośrednich

  Do wykorzystania na koszty pośrednio związane z projektem, **w tym** koszty udostępnienia publikacji i/lub danych badawczych w otwartym dostępie.

Wdrażanie i raportowanie Planów Zarządzania Danymi

Wdrażanie i raportowanie Planu Zarządzania Danymi

- **Plan zarządzania danymi może zmieniać się w trakcie realizacji projektu (zalecane)**
  Nie ma obowiązku informowania NCN o zmianach.

- **W raporcie rocznym** należy opisać kwestie związane z udostępnianiem danych powiązanych z publikacjami.
- **W raporcie końcowym** należy opisać stan faktyczny na koniec realizacji projektu: *planowano vs. zrealizowano.*

Kontakt:

sz.kubik@uj.edu.pl
jakub.rusakow@gumed.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.

Narodowe Centrum Nauki
Zespół ds. Otwartej Nauki
otwarta.nauka@ncn.gov.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.