DATA MANAGEMENT PLAN

# 1. Data description and collection or re-use of existing data

## How will new data be collected or produced and/or how will existing data be re-used?

Projekt wykorzystuje głównie wtórne źródła danych z publicznie dostępnych repozytoriów i dokumentów strategicznych, w szczególności z OECD AI Policy Observatory w formie eksportu CSV i plików PDF. Dane jakościowe będą pozyskane z transkrypcji wywiadów eksperckich przeprowadzonych w ramach działania. Nagrania audio będą digitalizowane do formatu MP3. Pochodzenie danych będzie dokumentowane poprzez systematyczne oznaczanie źródeł, dat pozyskania i wersji. W projekcie wykorzystywane będą też standardy BERTopic w wersji 1.16, Python 3.9+, HDBSCAN oraz UMAP.

## What data (for example the types, formats, and volumes) will be collected or produced?

Dane pierwotne obejmują korpus dokumentów strategicznych AI składający się z około dokumentów w formatach.

PDF i CSV o łącznej wielkości około GB. Ponadto zostanie utworzonych transkrypcji wywiadów eksperckich w formatach TXT i DOCX o łącznej objętości około stron. Nagrania audio wywiadów będą przechowywane w formacie.

MP (około godzin) i usuwane po zakończeniu transkrypcji.

Dane przetworzone będą składać się z wyników analiz BERTopic obejmujących około tematów w formatach CSV i.

JSON o wielkości około 100MB oraz analizy wywiadów w formatach DOCX/PDF.. Metadane instrumentów polityki publicznej będą przechowywane w formatach CSV i Excel o wielkości około 10MB. Mapa klastrów krajów będzie dostępna w formatach CSV i JSON o wielkości około 5MB. Słownik pojęć otwartej innowacji zostanie utworzony w formatach DOCX i PDF o objętości około stron. Formatami priorytetowymi są CSV, JSON i TXT jako formaty otwarte oraz DOCX i PDF dla dokumentacji.

• Dokumentacja i jakość danych.

# 2. Documentation and data quality

## What metadata and documentation (for example methodology or data collection and way of organising data) will accompany data?

Metadane będą zawierać identyfikator pliku i wersję każdego dokumentu. Źródło pochodzenia będzie oznaczone poprzez URL OECD lub kod wywiadu. Każdy plik będzie posiadał datę pozyskania lub utworzenia. Autor lub osoba odpowiedzialna będzie identyfikowana poprzez ORCID 0000-0003-1181-669X. Metodologia przetwarzania będzie opisana poprzez parametry BERTopic i HDBSCAN. Metadane będą zawierać definicje zmiennych i jednostki pomiarowe oraz format i kodowanie pliku.

Organizacja danych będzie oparta na zdefiniowanej strukturze folderów obejmującej katalogi raw\_data, processed\_data, results i documentation. Książka kodowa w formacie Excel będzie zawierać definicje wszystkich zmiennych. Standardem metadanych dla repozytoriów będzie Dublin Core.

## What data quality control measures will be used?

Kontrola jakości będzie obejmować weryfikację kompletności danych z OECD poprzez automatyczne sprawdzenie brakujących pól. Transkrypcje będą walidowane poprzez kontrolę 20% materiału. Parametry BERTopic będą kalibrowane poprzez testy z różnymi ustawieniami. Wyniki będą porównywane z literaturą przedmiotu w ramach triangulacji metodologicznej. Wszystkie wersje plików będą oznaczane czasowo. Błędy systematyczne będą eliminowane poprzez standaryzację procedur. Wszystkie decyzje metodologiczne będą dokumentowane. Dane ilościowe będą czyszczone z wartości odstających.

# 3. Storage and backup during the research process

## How will data and metadata be stored and backed up during the research process?

Głównym repozytorium będzie zabezpieczony serwer uczelniany Politechniki Gdańskiej. Pojemność 10GB przestrzeni dyskowej będzie wystarczająca dla całego projektu. Struktura będzie obejmować dedykowane katalogi dla każdego typu danych.

Kopie zapasowe będą tworzone automatycznie codziennie na szyfrowanym dysku zewnętrznym. Kopie tygodniowe będą przechowywane w chmurze uczelnianej. Kopie miesięczne będą tworzone na drugim dysku zewnętrznym przechowywanym w innej lokalizacji. Odpowiedzialnością dr inż. Marita Mcphillips będzie zarządzanie głównym systemem, natomiast wsparcie IT uczelni będzie odpowiedzialne za infrastrukturę serwerową.

OSF, Wniosek o przyznanie środków finansowych na realizację działania naukowego Strona ID: 655050, 2025-07- 13:28:

Dane jakościowe będą szyfrowane. Dostęp do wszystkich danych będzie ograniczony wyłącznie do osoby realizującej działanie poprzez autoryzację dwuskładnikową. Nagrania wywiadów będą automatycznie usuwane po zakończeniu transkrypcji.

• Wymogi prawne, kodeks postępowania.

## How will data security and protection of sensitive data be taken care of during the research?

# 4. Legal requirements, codes of conduct

## If personal data are processed, how will compliance with legislation on personal data and on data security be ensured?

Wywiady będą anonimizowane poprzez usunięcie danych identyfikujących. Respondenci wyrażą świadomą zgodę na udział w badaniu. Przetwarzanie danych będzie zgodne z polityką ochrony danych Politechniki Gdańskiej oraz wymogami RODO.

## How will other legal issues, such as intelectual property rights and ownership, be managed? What legislation is applicable?

Obowiązujące przepisy.

Wykorzystywane dane z OECD AI Policy Observatory są publicznie dostępne na licencji otwartej. Transkrypcje wywiadów będą własnością intelektualną projektu. Wszystkie procedury będą zgodne z założeniami etyki projektów NCN.

• Udostępnianie i długotrwałe przechowywanie danych.

# 5. Data sharing and long-term preservation

## How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons?

Zanonimizowany korpus danych z adnotacjami otwartej innowacji oraz mapa polityk AI zostaną udostępnione publicznie w repozytorium MOST Wiedzy Politechniki Gdańskiej najpóźniej w ciągu miesięcy od zakończenia projektu pod licencją Creative Commons BY 4.0.

## How will data for preservation be selected, and where will data be preserved long-term (for example a data repository or archive)?

repozytorium lub archiwum danych).

Do długotrwałego przechowywania przeznaczone będą korpus dokumentów z metadanymi, wyniki analiz BERTopic oraz zanonimizowane transkrypcje. Dane będą zarchiwizowane w repozytorium MOST Wiedzy Politechniki Gdańskiej.

## What methods or software tools will be needed to access and use the data?

Dostęp do danych będzie zapewniony poprzez interfejs repozytorium MOST Wiedzy umożliwiający przeglądanie i pobieranie (formaty CSV/JSON). Standardowymi narzędziami będą Python z bibliotekami pandas i numpy oraz Excel.

Oprogramowanie do konwersji nie będzie wymagane dla danych w formatach otwartych. Narzędzia BERTopic będą dostępne jako pakiet Python w formie open source.

Mechanizm udostępniania będzie obejmować bezpośrednie pobieranie z repozytorium. Na żądanie będzie możliwa wysyłka większych zestawów danych poprzez e-mail. Długoterminowa dostępność będzie zapewniona przez minimum.

lat zgodnie ze standardami NCN.

## How will the application of a unique and persistent identifier (such us a Digital Object Identifier (DOI)) to each data set be ensured?

Każdy zestaw danych otrzyma unikalny identyfikator DOI poprzez system MOST Wiedzy Politechniki Gdańskiej. Dane będą zatem oznaczone stałymi identyfikatorami zgodnie ze standardami FAIR.

• Zadania związane z zarządzaniem danymi oraz zasoby.

# 6. Data management responsibilities and resources

## Who (for example role, position, and institution) will be responsible for data management (i.e the data steward)?

a)włączenia go do planu zadaniowo-finansowego podmiotu;.

b)zatrudnianiawykonawcówzbiorowychniezbędnychdorealizacjidziałanianaukowegonapodstawieuzgodnionej z wykonawcami formy zatrudnienia (umowa o pracę, umowa o dzieło, umowa zlecenie); c)zapewnienia warunkówdorealizacjidziałanianaukowego, wtymudostępnienia przestrzeni biurowej/laboratoryjnej oraz aparatury naukowo-badawczej niezbędnej do jego realizacji; d)zapewnienie obsługi administracyjno-finansowej realizacji działania naukowego; e)sprawowanianadzorunadrealizacjądziałanianaukowegoiprawidłowością wydatkowanych natencelśrodków finansowych;.

7. zapoznałem/am się z zasadami doręczania decyzji Dyrektora NCN;.

8. wyrażamzgodęnadokonanieweryfikacjiwnioskuprzypomocyoprogramowania antyplagiatowego oraz umieszczenie treści wniosku w bazie danych oprogramowania;.

9. zapoznałem/am sięztreścią Kodeksu Narodowego Centrum Naukidotyczącegorzetelnościbadańnaukowychi starania o fundusze na badania i zobowiązuję się do jego stosowania;.

10.świadomy/aodpowiedzialności prawnejwynikającejzprzekazanianieprawdziwych informacjizapewniam,że informacjezawarteweniniejszymwnioskuofinansowaniedziałanianaukowegoorazdokumentach doniego dołączonychzłożonymzapośrednictwem systemu OSF(Obsługa Strumieni Finansowania), sązgodnezestanem faktycznym i prawnym;.

11. akceptujęogólnewarunkiumowynafinansowanieirealizacjędziałanianaukowego(treśćogólnychwarunków umowy);.

12. jestemświadomy/a,że–wprzypadkuzakwalifikowania wnioskudofinansowania–dniemrozpoczęciarealizacji działanianaukowegojestdzień, wktórymdecyzja Dyrektora Narodowego Centrum Naukiprzyznającafinansowanie stała się ostateczna;.

13. podmiot, któryreprezentuję, niepozostajepodzarządemkomisarycznym aninieznajdujesięwtokulikwidacjilub postępowania upadłościowego.

OSF, Wniosek o przyznanie środków finansowych na realizację działania naukowego Strona ID: 655050, 2025-07- 13:28:

OCHRONA DANYCH OSOBOWYCH.

INFORMACJA O ZASADACH PRZETWARZANIA DANYCH OSOBOWYCH.

Administratorem Pani/Pana danych osobowych jest Narodowe Centrum Nauki z siedzibą w Krakowie przy ul.

Twardowskiego 16, 30- Kraków.

Kontakt do Inspektora Ochrony Danych: iod@ncn. gov. pl. Pani/Pana dane będą przetwarzane w celach:

dokonania oceny wniosku o finansowanie działania naukowego, a.

nadzoru, obsługi finansowo-księgowej, kontroli w trakcie jak i po zakończeniu działania naukowego, oceny jegob.

realizacji i rozliczenia umów o finansowanie, przeprowadzania ewaluacji realizacji zadań Centrum, sprawozdawczości, upowszechniania w środowiskuc. naukowym informacji o ogłaszanych przez Centrum konkursach, realizacji innych czynności regulowanych przepisami prawa powszechnie obowiązującego oraz w celach archiwalnych.

OSF, Wniosek o przyznanie środków finansowych na realizację działania naukowego Strona ID: 655050, 2025-07- 13:28:

## What resources (for example financial and time) will be dedicated to data management and ensuring the data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?