Zarządzanie danymi badawczymi w językoznawstwie

dr Agnieszka Dziob-Zadworna[1,2], dr Jan Wieczorek[1]
Politechnika Wrocławska / CLARIN-PL

[1] Wydział Informatyki i Telekomunikacji / Katedra Sztucznej Inteligencji
[2] Wydział Zarządzania, Katedra Systemów Zarządzania i Organizacji

NARODOWE CENTRUM

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.

Ministerstwo Edukacji i Nauki

Dlaczego zarządzanie danymi badawczymi jest ważne w językoznawstwie (1)

1. Dane badawcze to clue cyfrowej humanistyki - bez danych nie ma badań w tym paradygmacie
2. EKONOMIA: Wytworzenie rzetelnych danych jest kosztowne - angażuje środki finansowy i czas specjalistów. Dzięki sprawnemu zarządzaniu możemy z nich korzystać wielokrotnie
3. BEZPIECZEŃSTWO third parties: Niektóre dane mogą być z różnych powodów chronione (materiał dowodowy, dane niejawne, bezpieczeństwo osób wywiadowanych)
4. BEZPIECZEŃSTWO danych: żeby nie zginęły w wyniku rozmaitych wydarzeń losowych
5. WIDOCZNOŚĆ: dane muszą być wyszukiwalne i logicznie deponowane (bo inaczej realnie ich nie ma); zwiększamy widoczność, co przekłada się na większy wpływa na naukę i większe szanse na cytowanie;
6. WERYFIKOWALNOŚĆ: ułatwienie procesu weryfikowania danych naukowych; zwiększenie zaufania do badań naukowych;

Dlaczego zarządzanie danymi badawczymi jest ważne w językoznawstwie (2)

1. KLAROWNOŚĆ: żeby podatnik wiedział, że pieniądze nie poszły na marne; by istniała społeczna kontrola wydatków
2. NAUKA OBYWATELSKA: polityka naukowe UE wskazuje, że Citizen Science jest jednym z ośmiu filarów przyszłości nauki - CS musi mieć dostęp do danych wytworzonych również przez instytucjonalną naukę
3. WYDAJNOŚĆ: kiedy wiemy, czym już dysponujemy, możemy racjonalnie wydawać środki (istotne dla podatnika, kierownictwa jednostek oraz naukowców, którzy wiedzą, czego brakuje)
4. WZGLĘDY FORMALNE: spełnienie wymogów instytucji finansujących badania

Charakterystyka danych w językoznawstwie (1)

- językoznawcy są świetnymi biorcami danych (prawie wszystkie dane tekstowe zbierane przez innych naukowców mogą być przydatne)
- dane tekstowe są indeksowane przez wiele wyszukiwarek (+ duży wybór, optymalne pokrycie; - czasami trzeba korzystać z kilku wyszukiwarek)
- danych tekstowych jest już dużo (choć ich dystrybucja jest nierówna, polszczyzna do niedawna była kwalifikowana jako “underresourced”)
- językoznawców nie bolą dane zanonimizowane - interesuje nas głównie forma i struktura
- produkujemy dane tekstowe, które mogą być bardzo użyteczne dla przedstawicieli innych dyscyplin (np. zdigitalizowane listy urzędników osiemnastowiecznych mogą być istotne dla historyka języka, ale również dla historyka ekonomii lub badacza procesów politycznych)

Charakterystyka danych w językoznawstwie (2)

- korpusy (np. NKJP, KPWr, Italian Drama Corpus, DiaBiz.Com, KorBa)
- korpusy treningowe
- wordnety: Słowosieć (plWordNet), African Wordnet (AWN), Czech WordNet
- tezaurusy (MeSH - tezaurus medyczny, Digizaurus - tezaurus sztuk pięknych, EuroVoc - tezaurus Unii Europejskiej)
- słowniki (Wielki Słownik Języka Polskiego)
- ontologie ogólne (SUMO, YAGO, DBPedia)
- ontologie dziedzinowe (np. Cambia - waluty, Harmonize - obsługa turystyki)
- opisy gramatyczne (drzewa zależności składniowych)
- modele językowe
- nagrania tekstów mówionych (np. Paralela)
- transkrypcje nagrań tekstów mówionych (np. Paralela)

Zasady FAIR

Findable: odnajdywalne

- dane muszą być łatwe do znalezienia (PID, bogaty opis metadanymi, metadane/dane umieszczone w przeszukiwalnym, indeksowanym repozytorium)

Accessible: dostępne

- droga dostępu do danych musi być jasno i klarownie opisana (o ile to możliwe - dane powinny być dostępne od ręki)

Interoperable: interoperacyjne

- dane są zdeponowane w ustandaryzowanych formatach umożliwiających przetwarzanie lub integrację (np. OWL, JSON, CSV, TEI itp.)

Reusable: możliwe do ponownego wykorzystania

- precyzyjny opis metadanymi, podanie informacji o licencji znacznie ułatwi późniejsze ponowne użycie naszych danych (i będzie wiadomo, komu przypisać zasługę)

Zasady otwierania danych badawczych

Dane powinny być tak otwarte, jak to możliwe i na tyle zamknięte, na ile to jest konieczne [*as open as possible, as closed as necessary*].

- **Maksymalne otwarcie danych ułatwi dostęp do nich i zwiększy ich używalność**
- **Jesteśmy ograniczeni licencjami, prawem autorskim, chrakterystyką danych (jak bardzo są wrażliwe), RODO**

Przykłady:

- **Korpus Listów Pożegnalnych (Monika Zaśko-Zielińska) - dostępne metadane/dane za zgodą**
- **PolitOrator (Rafał Zimny) - dostępne metadane/dane za zgodą**
- **Korpus Politechniki Wrocławskiej - dostępne metadane oraz dane (od ręki)**

Otwieranie danych badawczych - jak wybrać repozytorium (1)

Należy wybrać odpowiednie repozytorium:

Opcja 1: repozytorium uczelni / repozytorium ogólne

+ **wsparcie personelu uczelni**
+ **“bliskość”**
+ **spełnione są podstawowe warunki otwartości**
+ **ad maiorem academiae gloriam**

- **indeksowanie w wyszukiwarkach danych?**
- **bezpieczeństwo danych?**
- **uniwersalny standard opisu metadanymi?**
- **certyfikaty?**
- **obsługa licencjonowania?**

Nie zawsze FAIR

Otwieranie danych badawczych - jak wybrać repozytorium (2)

Należy wybrać odpowiednie repozytorium:

Opcja 2: repozytorium dziedzinowe

+ **przejrzyste certyfikowanie/bezpieczeństwo**
+ **indeksowane w wyszukiwarkach danych i metadanych**
+ **przejrzyste licencje**
+ **przejrzyste i standardowe metadane**
+ **informacja o dostępie i własności danych**
+ **niemal zawsze honorowane przez grantodawców**

- **trzeba wyszukać odpowiednie repozytorium**
- **wsparcie zazwyczaj tylko zdalne**
- **fumy akademii**
- **certyfikaty?**
- **obsługa licencjonowania?**

Zazwyczaj FAIR

Otwieranie danych badawczych - jak wybrać repozytorium (3)

Gwarancja FAIR - SSHOC

https://www.sshopencloud.eu/

SSHOC
social sciences & humanities open cloud

Partnerzy SSHOC to zarówno rozwijające się, jak i w pełni rozwinięte europejskie infrastruktury badawcze z dziedziny nauk społecznych i humanistycznych, a także stowarzyszenie europejskich bibliotek badawczych (LIBER). Partnerzy ci dysponują wiedzą specjalistyczną w zakresie całego cyklu danych, od ich tworzenia i przechowywania po optymalne ponowne wykorzystanie, szkolenia i propagowanie.

CEL: Zmiana dyscyplinowych jednostek i oddzielnych ośrodków w zintegrowaną, opartą na chmurze sieć połączonych ze sobą infrastruktur danych

Otwieranie danych badawczych - jak wybrać repozytorium (4)

https://www.sshopencloud.eu/data-catalogues

Virtual Language Observatory

dostawcy danych: https://vlo.clarin.eu/contributors

VLO – Virtual Language Observatory (CLARIN)

!TUTAJ!

Plan Zarządzania Danymi
Ogólne Informacje

Dlaczego Plan Zarządzania Danymi jest ważny w językoznawstwie?

- daje poczucie bezpieczeństwa, oferując standardy zabezpieczeń przed nieuprawnionym dostępem oraz regulując kwestię tworzenia kopii zapasowych;
  Przykład: certyfikat CoreTrustSeal

- umożliwia wyprodukowanie danych dobrej jakości i zarządzanie nimi przez cały okres procesu badawczego bez utraty tej jakości;
- kładzie nacisk na metadane i dokumentację;
- reguluje kwestię kontroli jakości danych
  Przykład: dokumentacja dla anotatorów, która umożliwia tworzenie danych strukturyzowanych oraz wymusza kontrolę jej jakości (np. metodą 2+1 z pomiarem zgodności) oraz określa, jakie metadane będą zapisywane;

Dlaczego Plan Zarządzania Danymi jest ważny w językoznawstwie?

- ułatwia współpracę badawczą dzięki standaryzacji formatów i polityce dostępności do danych;
- wspiera pozyskiwanie danych do celów badawczych i ich re-używalność w innych procesach badawczych;
- umożliwia łączenie danych w sieci światowych połączonych danych (przykład: Linked Open Data);
- do tworzenia planu zarządzania danymi i upubliczniania go może być wykorzystany DMPTool (https://dmptool.org/); na podanej stronie znajdują się gotowe DMP;

Przykład: tezaurus dziedzinowy, który był tworzony od podstaw, z wykorzystaniem relacji RDF (standardu W3C), co umożliwia jego połączenie z siecią światowych danych; tezaurus od początku stanowił dobrej jakości dane badawcze i był rozbudowywany z wykorzystaniem schematu danych; zostanie udostępniony na licencji Creative Commons z możliwością do przeglądania i do pobrania

Dlaczego Plan Zarządzania Danymi jest ważny w językoznawstwie?

- umożliwia długotrwałe przechowywanie danych na repozytoriach;
- umożliwia dodawanie nowych wersji zbiorów danych;
- umożliwia ich wyszukiwanie i przeszukiwanie dzięki identyfikatorom;

Przykład: na platformie https://clarin-pl.eu/dspace/ znajduje się korpus Użytkownika, który, dzięki standardowi opisu CMDI, wymaganemu na platformie, zyskuje opis metadanymi, licencję oraz identyfikator handle.net, dzięki czemu jest możliwy do wyszukania przez innych Użytkowników, którzy chcą z niego korzystać; właściciel korpusu postanowił wgrać jego kolejną wersję, która również uzyskuje opis metadanymi CMDI oraz identykor handle.net; platforma DSpace jest połączona z centralnymi repozytoriami CLARIN ERIC, dlatego możliwość wykorzystania korpusu nie ogranicza się do Polski.

Opis danych oraz pozyskiwanie lub ponowne wykorzystanie dostępnych danych

Opis danych (1)

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

Dane pierwotne (nowe):

- **typ danych** - słownik/ontologia/korpus/transkrypcja/nagranie wywiadu…
- **sposób pozyskania danych** - digitalizacja, kwerenda, nagranie, transkrybowanie…
- **podmiot wytwarzający** - jednostka wytwarzające dane, osoba odpowiedzialna, współtwórcy…
- **zakres** - które elementy wytworzonych danych chcemy/możemy udostępnić (a może tylko metadane)
- **licencja** - jaka licencja będzie obowiązywać użytkownika naszych danych
- **środki, które posłużyły do wytworzenia danych** - sprzęt (np. nagranie za pomocą specjalistycznych mikrofonów), oprogramowanie:
  - narzędzia do OCR
  - narzędzia do transkrypcji (np. Transkribus)
  - narzędzia korpusowe (np. Korpusomat, SketchEngine, NoSketchEngine)
  - narzędzia anotacyjne (np. DoccAno, Inforex, Atlas)
  - narzędzia przetwarzania mowy (np. mowa.clarin-pl.eu)
  - narzędzia leksykograficzne (np. WordNetLoom)
  - infrastruktura naukowa wykorzystana przy wytworzeniu (np. CLARIN-PL, DARIAH-PL)

Opis danych (2)

Dane wtórne (istniejące wcześniej):

- **własność danych** - kto jest właścicielem danych wtórnych (to wynika z licencji lub umowy)
- **sposób udokumentowania pochodzenia danych** - w DMP oraz w metadanych powinniśmy umieścić informację o źródle
- **licencja źródłowa**
- **kontrola jakości i spójności danych**
- **charakter danych źródłowych** (czy efekt digitalizacji, transkrypcja itp.)

Opis danych

Jakie dane będą pozyskiwane lub wytwarzane w projekcie?

- **rodzaj gromadzonych danych** (np. nagrania, leksykony, teksty, materiał graficzny, drzewa zależnościowe)
- **w jakim formacie będą używane i udostępniane** (formaty otwarte: csv, odt, rtf, txt, html, xml, png, flac; zamknięte: doc, docx, xls, mp3)
- **format/oprogramowanie** - format zamknięty bywa użyteczny, ale zmusza nas do wykorzystania i zakupu konkretnych programów komputerowych
- **szacunkowa objętość danych** - w odniesieniu do rodzaju gromadzonych danych (lekkie - tekst, ciężkie - nagrania bezstratne)

Dokumentacja i jakość danych

Dokumentacja i jakość

Metadane i dokumentacja dot. danych w językoznawstwie

CMDI

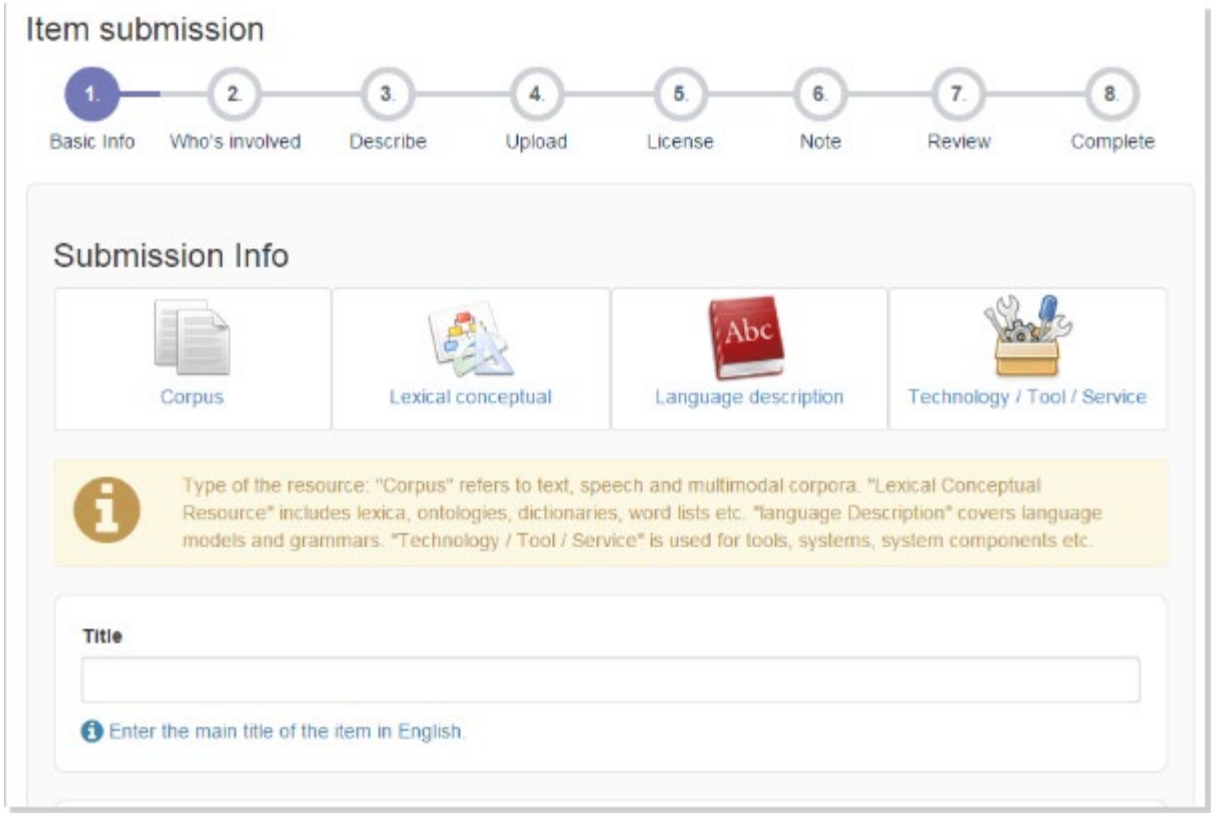
Standard metadanych CMDI (Component Metadata Infrastructure) jest zgodny z normami ISO i polityką CLARIN ERIC

TEI

Standard TEI (Text Encoding Initiative) jest coraz częściej wykorzystywany w naukach humanistycznych i społecznych; pozwala na odtworzenie warstwy wizualnej tekstu, która jest zapisywana w metadanych; strona standardu: https://tei-c.org/.

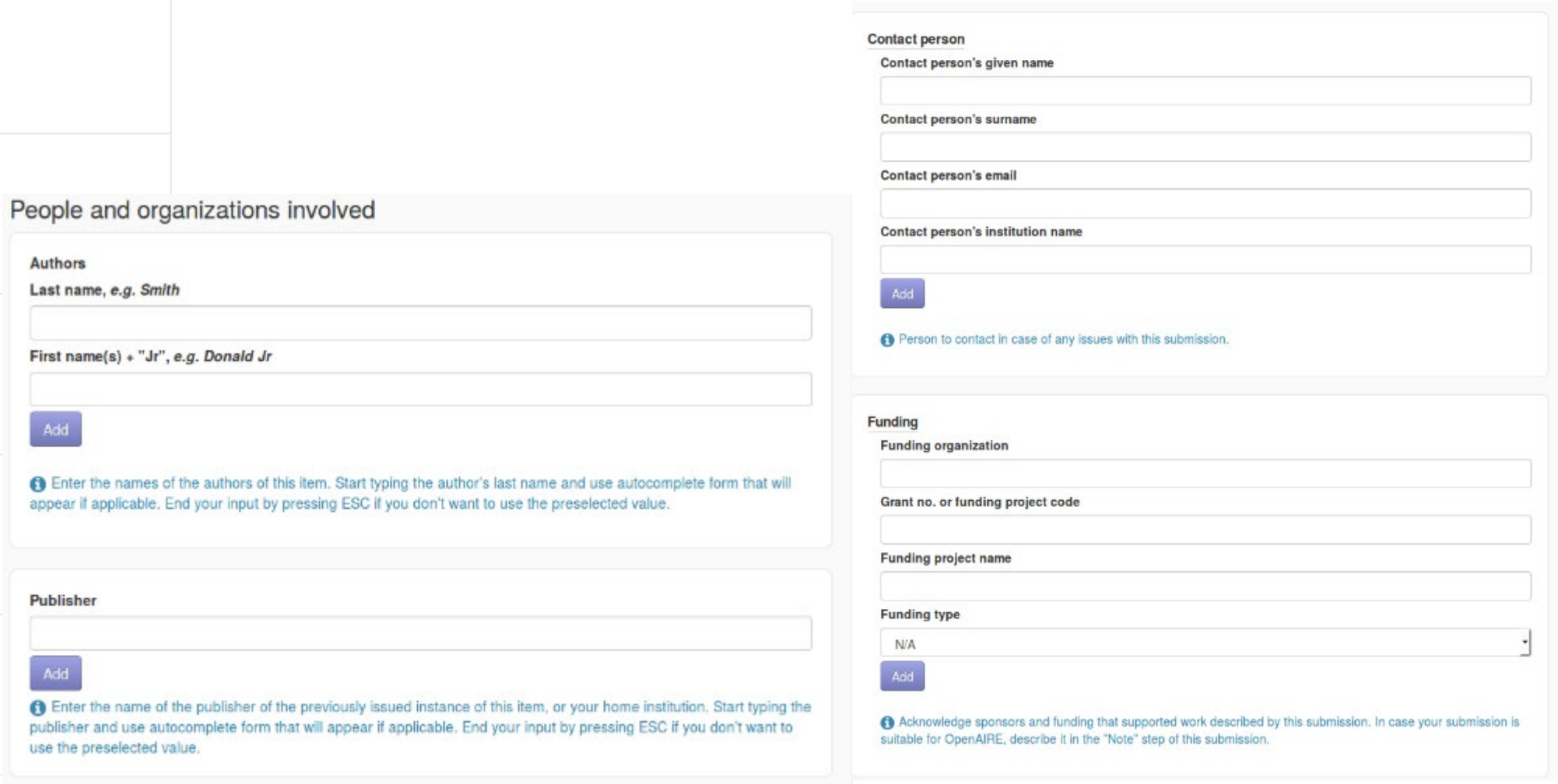
Dokumentacja i jakość

1. Typ danych



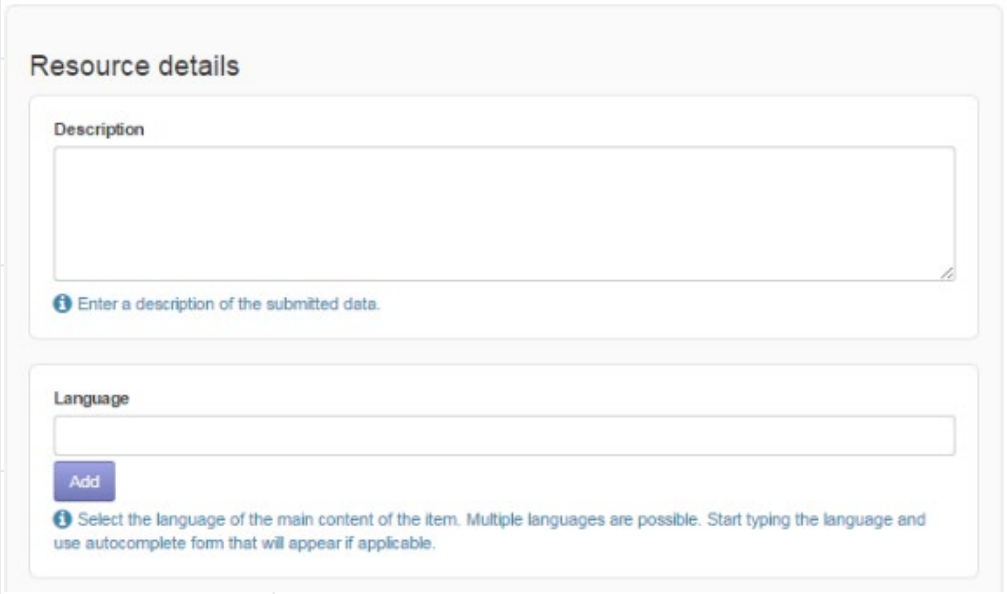
Dokumentacja i jakość

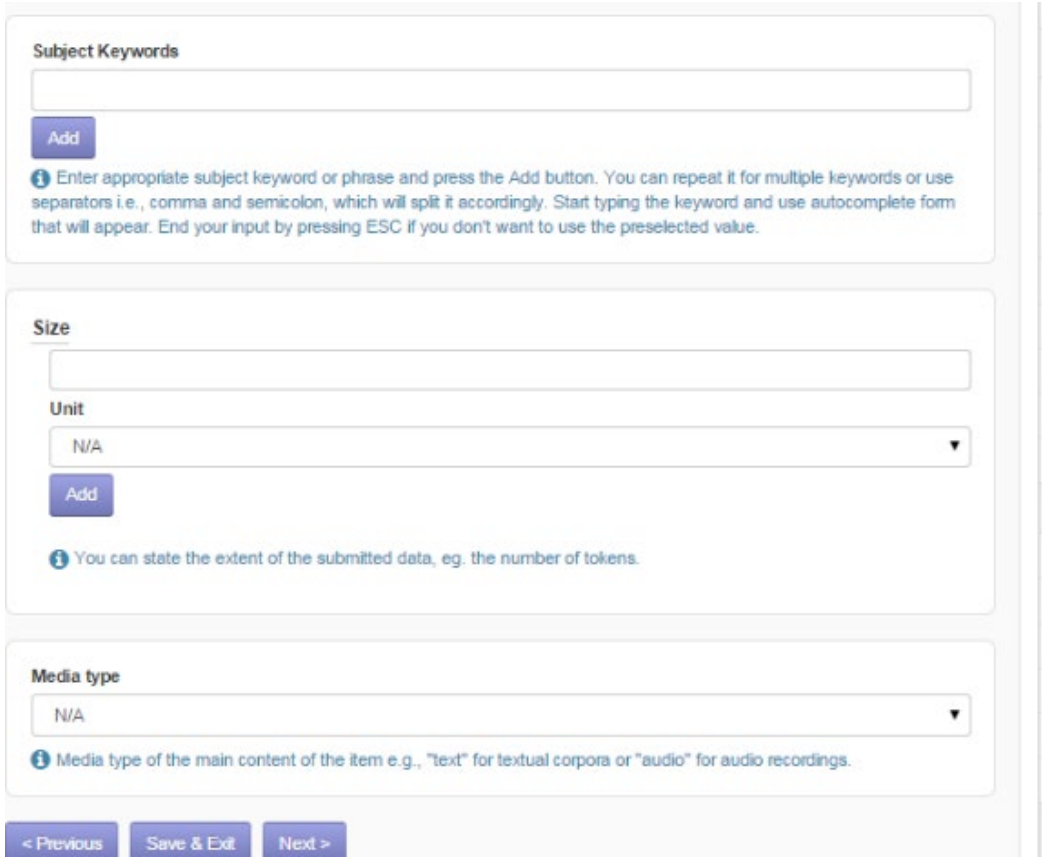
2. Opis zasobu



Dokumentacja i jakość

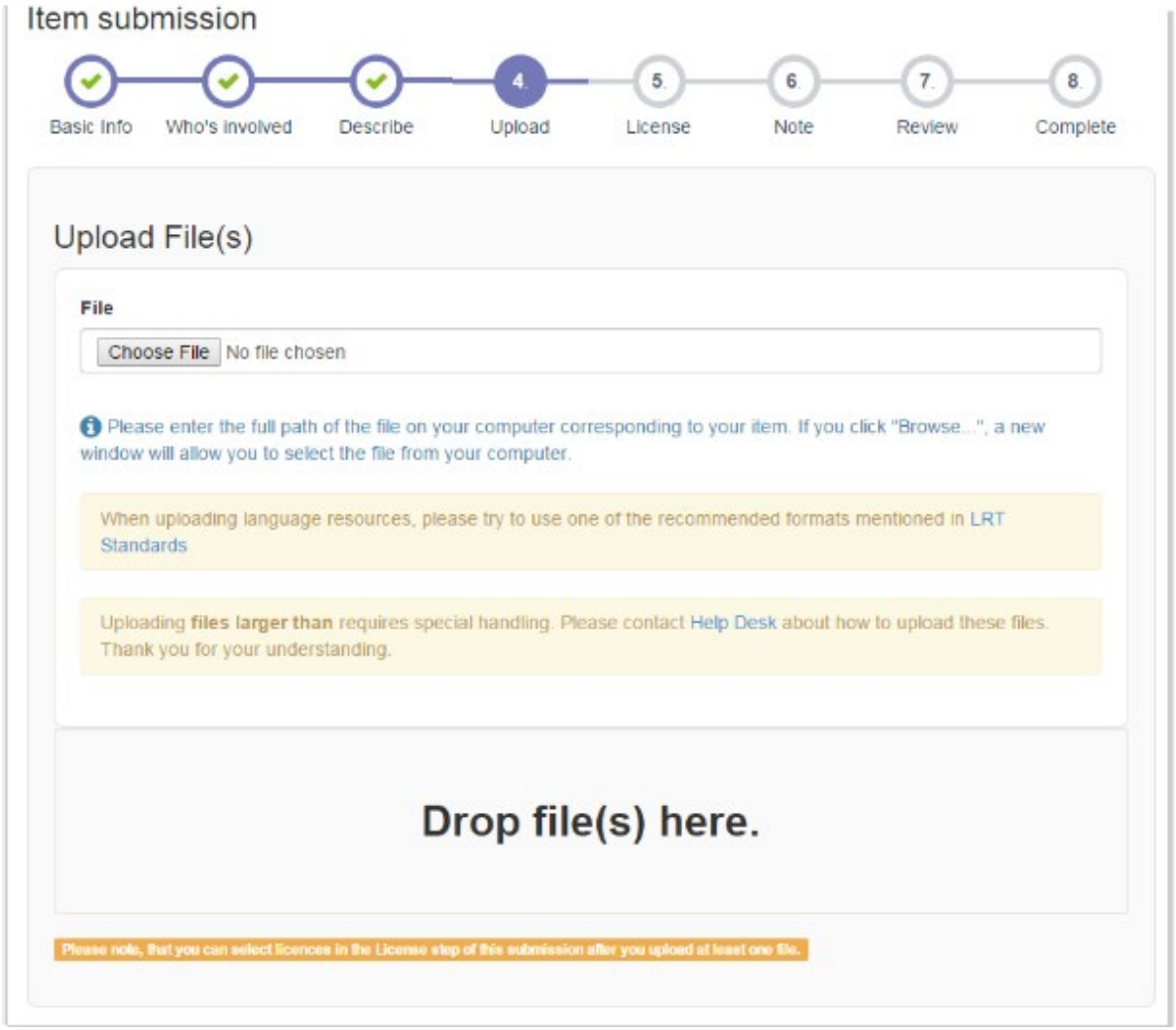
2. Opis zasobu





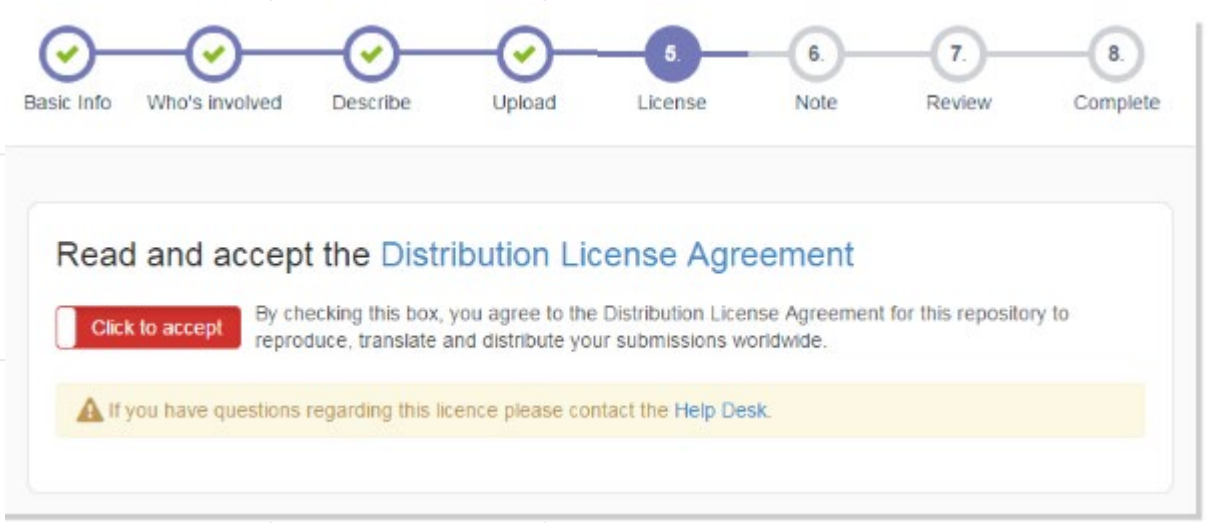
Dokumentacja i jakość

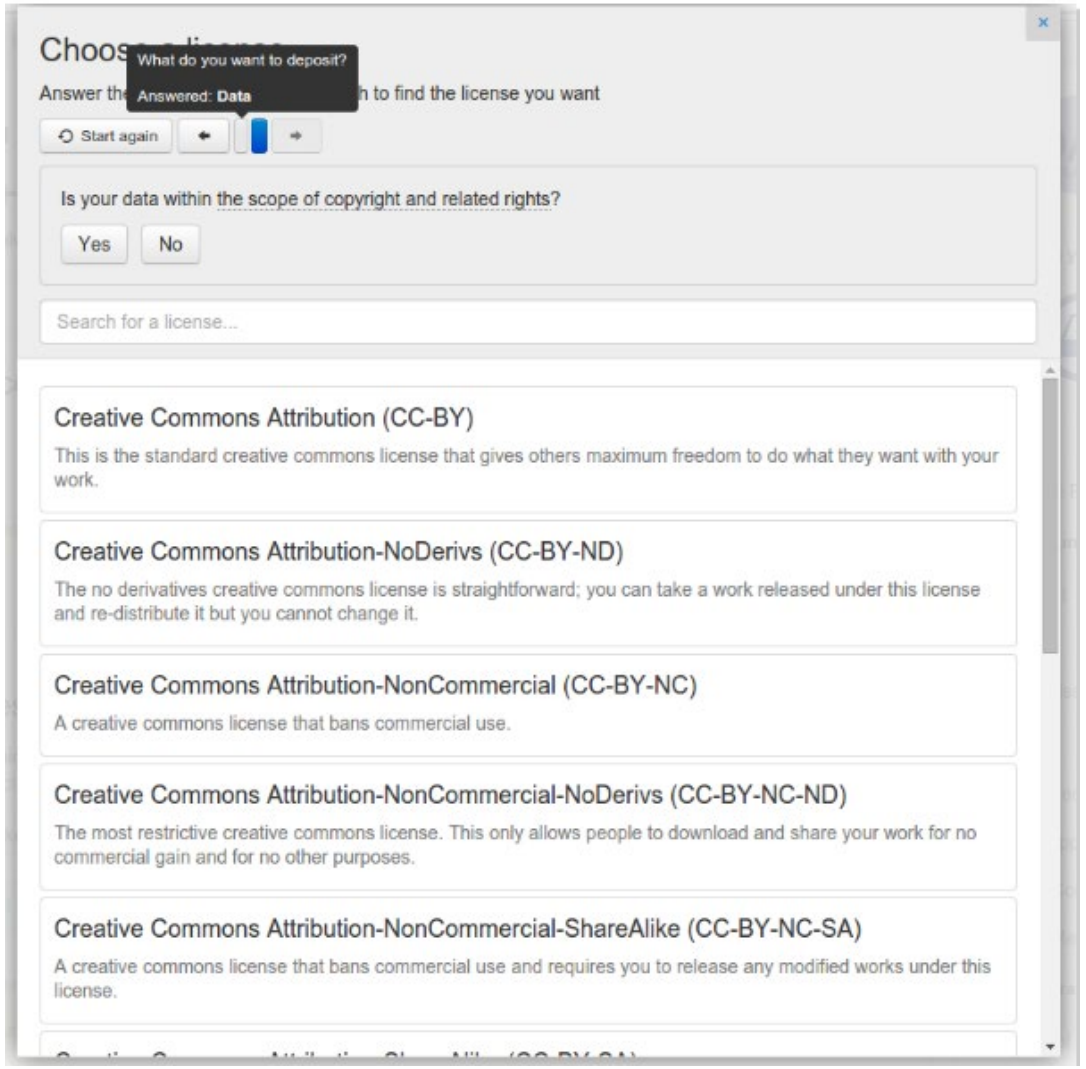
3. Ładowanie plików



Dokumentacja i jakość

4. Wybór licencji





Dokumentacja i jakość

Pozyskiwanie i przetwarzanie danych:

- stara dyscyplina:
  - duży udział danych historycznych;
  - brak standardów danych (istnieją dopiero od bardzo niedawna, biorąc pod uwagę wiek dyscypliny!);
  - różne formaty (np. fiszki ręcznie zapisywane, nagrania mowy na taśmach magnetofonowych, słowniki historyczne, wywiady spisywane itp.);
  - niska (w stosunku do nauk technicznych i przyrodniczych) świadomość obiegu danych, metod ich pozyskiwania, standardów opisu i licencji;
  - niska świadomość tego, że zasoby danych również podlegają cytowaniom i liczą się do dorobku badacza;
  - materiał trzymany w prywatnych archiwach badaczy.

Dokumentacja i jakość

Pozyskiwanie i przetwarzanie danych:

- dyscyplina humanistyczna:
  - duży nacisk na indywidualny kontakt badacza z materiałem (brak intersubiektywizmu badawczego);
  - przyzwyczajenie do badań jakościowych, prowadzonych metodami humanistycznymi (rozciągnięte czasowo analizy prowadzone na niewielkich próbkach danych);
  - tradycja prowadzenia badań w terenie, zaczerpnięta z nauk społecznych (okres przed upowszechnieniem się komputerów);
  - brak wypracowanych metod eliminowania błędów w interpretacji danych (wiąże się z indywidualizmem badawczym);
  - przewaga metod jakościowych nad ilościowymi (te drugie reprezentowane przez lingwistykę kwantytatywną, która korzysta z metod matematycznych).

Przechowywanie i tworzenie kopii zapasowych podczas badań.

Przechowywanie i kopie zapasowe

Przechowywanie danych i metadanych

Certyfikat CoreTrustSeal (https://www.coretrustseal.org/) określa:

- częstotliwość wykonywania kopii zapasowych (ustawiona automatycznie, nawet codziennie)
  - zasada 3-2-1 - przechowywanie w trzech kopiach na co najmniej dwóch nośnikach, z czego jeden poza siedzibą instytucji;
- warunki techniczne serwerów głównego i serwerów zapasowych;
- warunki techniczne systemu repozytoryjnego i standardy jego obsługi;
- politykę postępowania kryzysowego w przypadku utraty danych lub ich uszkodzenia.

Źródło: CoreTrustSeal-Requirements-2023-2025_v01.00
https://doi.org/10.5281/zenodo.7051011

Przechowywanie i kopie zapasowe

Przechowywanie danych i metadanych

Systemy do gromadzenia danych i ich wersjonowania:

- S3 (MinIO) - umożliwia wersjonowanie automatyczne;
- GitLab - umożliwia wersjonowanie automatyczne i zarządzanie pracami (również na danych);
- NextCloud - głównie dla dokumentów roboczych;
- DSpace - dane wgrywane przez Użytkownika, wersjonowanie ręczne, do publikacji danych.

Systemy pozwalają na współdzielenie dokumentów, pracę zespołową, wgrywanie dokumentów z dowolnego miejsca.

Do badań niekomercyjnych są udostępniane bezpłatnie.

Ciekawostka: 31 marca to Światowy Dzień Backupu!

Przechowywanie i kopie zapasowe

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

Wymogiem Politechniki Wrocławskiej jest przechowywanie danych wrażliwych jedynie na repozytoriach PWr (fizycznie: Wrocławskie Centrum Sieciowo-Komputerowe), gdzie dostęp jest przez VPN lub z sieci wewnętrznej PWr za pomocą logowania Active Directory.

Sposoby postępowania z danymi wrażliwymi regulują umowy pomiędzy jednostkami.

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?

Nie zajmujemy się bezpośrednio przetwarzaniem danych osobowych, dyspozytorem jest Politechnika Wrocławska, która ma standardy w zakresie RODO.

Anonimizacja (usunięcie danych wrażliwych) lub pseudoanonimizacja (zastąpienie danych wrażliwych innymi, podobnymi):

- jest wymagana, jeśli dane będą dalej udostępniane (np. dla celów badawczych);
- jest przeprowadzana za pomocą preinstalowanego narzędzia (Anonimizator) instalowanego na komputerze użytkownika (lub, w przypadku “małych badań” w chmurze);
- dane przetworzone przez aplikację wymagają ponownego sprawdzenia;
- w tej chwili nie jest procesem odwracalnym (ale pracujemy nad tym).

Przykład: Użytkownik prowadzi prace badawczo-rozwojowe w zakresie automatycznego rozpoznania nazw własnych. Dysponuje danymi z urzędów, zawierającymi dane osobowe. Przed rozpoczęciem prac konieczna jest ich anonimizacja.

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?

Niektóre badania wymagają pozyskania zgody uczestników na udział (np. badania z zakresu interakcji człowiek-komputer, tzw. UX, tworzenie i przetwarzanie korpusów mowy). Zgoda musi zawierać informację **czy**, **komu, w jaki sposób** i **na jak długo** badacz będzie udostępniał dane. Jeśli dane mają być otwarte - zgoda również powinna zawierać taką informację, jak też informację o anonimizacji:

- w przypadku pełnej anonimizacji zgoda nie jest wymagana, ale
- dobrze ją pozyskać, ponieważ zgoda obejmuje informację, że uczestnik może ją wycofać,
- więc niepoinformowanie o anonimizacji i udostępnieniu zanonimizowanych danych, uczestnik może wycofać zgodę na udostępnianie.

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?

Przykład:

Administratorem danych osobowych jest Politechnika Wrocławska z siedzibą przy Wybrzeżu Wyspiańskiego 27 we Wrocławiu. Dotyczy to wszystkich sytuacji gdy uczelnia decyduje o celach i sposobach przetwarzania danych osobowych.

Administrator wyznaczył Inspektora Ochrony Danych Osobowych, z którym w sprawach ochrony danych osobowych możesz się kontaktować przez e-mail: iod@pwr.edu.pl.

W celu skontaktowania się z Administratorem Danych Osobowych w sprawach dotyczących przetwarzania i ochrony swoich danych skorzystaj z formularza kontaktowego [https://pwr.edu.pl/kontakt]

Szczegółowe informacje o tym jak Politechnika Wrocławska przetwarza dane osobowe znajdziesz w politykach prywatności [https://pwr.edu.pl/ochrona-danych-osobowych/polityki-prywatnosci]

Wyrażam zgodę na wykorzystanie mojego wizerunku w materiałach fotograficznych i filmowych powstałych w trakcie wydarzenia przez Politechnikę Wrocławską (PWr). Jednocześnie upoważniam PWr do decydowania o formie i czasie emisji mojego wizerunku zawartego w materiałach. Nie mam zastrzeżeń do wykorzystania filmu jako materiału zachęcającego społeczeństwo do korzystania z działań PWr. Wyrażenie zgody jest dobrowolne, ale niezbędne do wzięcia udziału w spotkaniu.

Wymogi prawne, kodeksy postępowania

Ustawa o prawie autorskim i prawach pokrewnych (tzw. prawo autorskie):

- podtyp prawa własności intelektualnej;
- przedmiot prawa autorskiego: utwór (np. publikacja, ale też korpus tekstów, narzędzie, model językowy);
- ogranicza możliwość dysponowania utworem bez zgody autora (również jego przetwarzania i przechowywania);
- wyjątki:
  - wykorzystanie w krytycznej opinii;
  - prawa gatunku (np. pastisz, karykatura);
  - nauczanie (np. podręczniki, czasopisma popularnonaukowe);
- tzw. prawo cytatu - wyjątek w prawie autorskim, pozwalający na wykorzystanie niewielkich fragmentów utworu, objętego prawami autorskimi, do własnych celów;

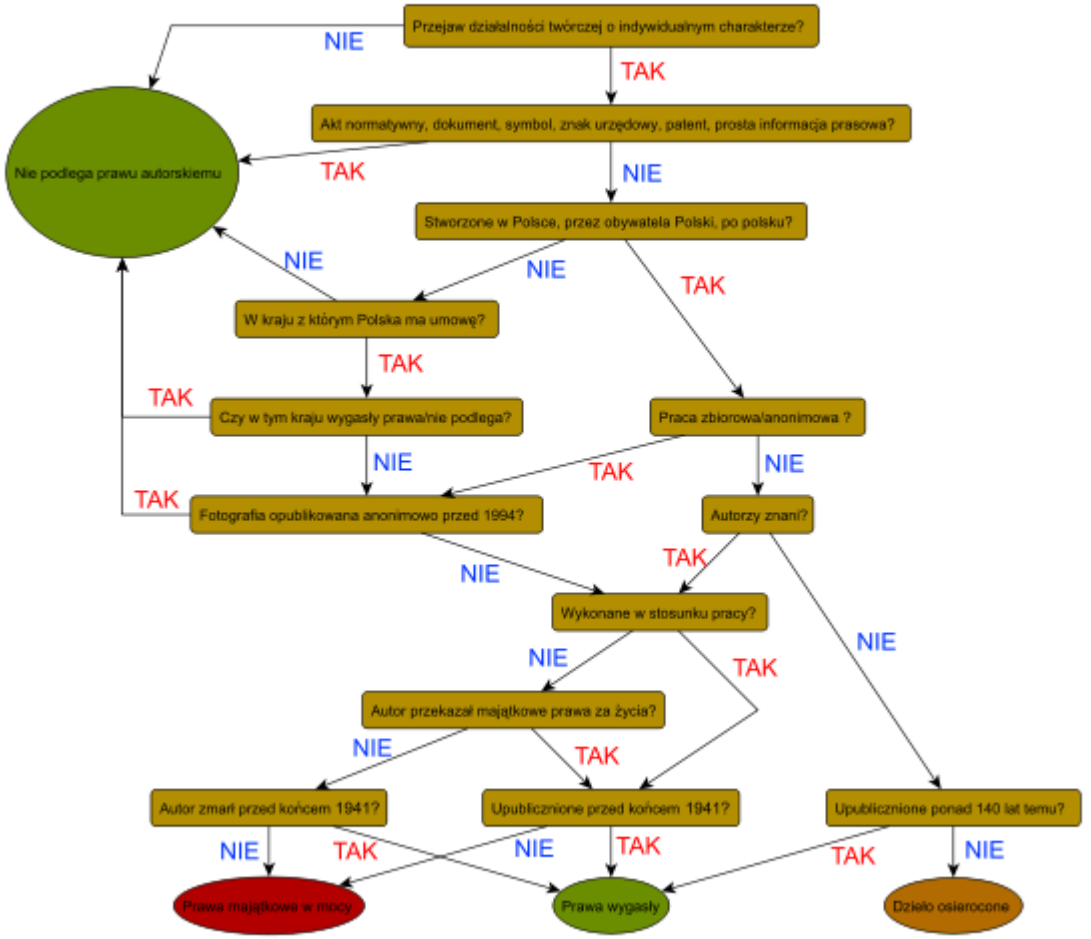
Prawo cytatu jest regulowane przepisami Unii Europejskiej (art. 6 ust. 2 lit b) i art. 9 dyrektywy 96/9/WE o ochronie baz danych i art. 5 ust. 3 lit d) dyrektywy 2001/29/WE).

Ze względu na pewną dowolność interpretacji przepisów staramy się korzystać z licencji Creative Commons i tekstów w domenie publicznej.

Wymogi prawne, kodeksy postępowania

Drzewo decyzyjne - Czy utwór podlega prawu autorskiemu?

Autor drzewa: dr Marek Maziarz



Wymogi prawne, kodeksy postępowania

Inne rodzaje regulacji:

- Większość konkursów, w których prace są finansowane ze środków publicznych (np. NCN, NCBR) wymaga stworzenia polityki zarządzania danymi badawczymi i udostępnienia (przynajmniej częściowo) danych na otwartych licencjach;
- Umowy dwustronne i konsorcyjne regulują kwestie zarządzania danymi, wykorzystywanymi w projektach badawczych;

Przykład: projekt finansowany z konkursu Szybka Ścieżka zakładał wytworzenie oprogramowania do odpowiadania na pytania zadane w języku naturalnym; wymogiem konkursu było, żeby część danych, wytworzonych w ramach projektu, udostępnić do potrzeb badawczych na otwartej licencji niekomercyjnej; na mocy porozumienia pomiędzy firmą a uczelnią zdecydowano się opublikować w otwartym dostępne słowniki i ontologie dziedzinowe, tworzone półautomatycznie.

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?

- **widoczność**: deponowanie materiału w indeksowanym repozytorium + publikacja z opisem zasobu
- **zakres**: metadane vs metadane + dane (publikacja pełna vs ograniczona)
- **czas**: czy dane będą udostępnione po zakończeniu projektu? w jego trakcie? na jak długo?
- **ograniczenia**: prawo, umowy, wrażliwość danych
- **powiązanie z publikacją wyników**: niektóre wydawnictwa (konferencje) wymagają opublikowania pełnych danych powiązanych z artykułem/książką/referatem
- **zgoda uczestników badań**: należy uzyskać taką zgodę (o ile dotyczy), w zgodzie powinna być informacja o tym, na jakiej licencji chcemy udostępnić dane

Udostępnianie i długotrwałe przechowywanie

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

Selekcja danych:

- **dane do utrwalenia**
  - **pełnego**
  - **ograniczonego**
- **dane do zniszczenia**

- dane utrwalone

Zasady FAIR

Nazewnictwo plików

Wersjonowanie

Identyfikator (PID)

Kryteria wyboru repozytorium

Typy repozytoriów

- repozytoria dziedzinowe vs ogólne
- repozytoria o wysokiej reputacji dla językoznawstwa
- Repozytoria posiadające certyfikat CoreTrust Seal

Wybór

- kwestie formalne wyboru (stosowanie FAIR, PID, licencja, czas przechowywania, embargo, rozmiar plików, inne)
- kwestie merytoryczne: „popularność” repozytorium wśród naukowców z dyscypliny, wpływa na widzialność, cytowalność, popularyzacja dorobku naukowego, indeksowanie

Rejestry repozytoriów

- Registry of Research Data Repositories [re3data.org]

Udostępnianie i długotrwałe przechowywanie

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

MECHANIZM:

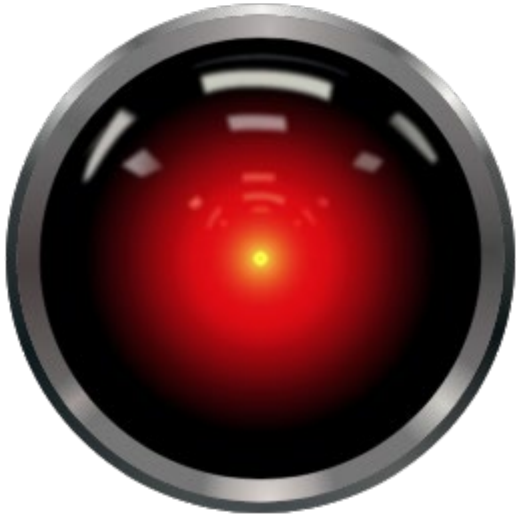
- **pełne otwarcie (repozytorium <- dane+metadane)**
- **ograniczone otwarcie (odpowiedź na żądanie, repozytorium <- metadane)**

REPOZYTORIUM:

- **konieczność zmiany formatu danych na format otwarty**

  (narzędzia CLARIN, wykorzystanie konwerterów)

Należy wykazać się pewną empatią w odniesieniu do przyszłych użytkowników



https://commons.wikimedia.org/wiki/File:HAL9000_-_Sharper_Reflections.svg

Udostępnianie i długotrwałe przechowywanie

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

Deponujmy dane w repozytoriach tworzących automatyczne identyfikatowy (PID).

SSHOC
social sciences & humanities open cloud

VLO – Virtual Language Observatory (CLARIN)

RepOD
Repozytorium Otwartych Danych

Zadania związane z zarządzaniem danymi oraz zasoby

Zadania związane z zarządzaniem danymi oraz zasoby

Role w projekcie:

- wytwórcy danych
  - Użytkownicy Infrastruktury;
  - zespół Lingwistów (dane strukturyzowane i niestrukturyzowane tworzone “na zamówienie innych zespołów”);
  - zespół Informatyków (duże dane, niestrukturyzowane, pobrane przez tzw. crawling lub ręcznie);
- konsumenci danych
  - Użytkownicy Infrastruktury;
  - zespół Lingwistów (strukturyzacja i normalizacja danych dostarczonych przez Informatyków);
  - zespół Informatyków (dane od Użytkowników i Lingwistów, dane strukturyzowane lub niestrukturyzowane);
- Zarządzanie danymi
  - Opiekun techniczny Infrastruktury;
  - Koordynatorzy współpracy z Użytkownikami (komercyjnymi i niekomercyjnymi).

Zadania związane z zarządzaniem danymi oraz zasoby

Przykład wykorzystania danych od Użytkowników

Zadanie polega na douczeniu algorytmu rozpoznawania jednostek nazewniczych (NER). Zespół Informatyków dostaje od Użytkowników dane, pochodzące z ich korpusów, które są złożone do systemu S3. Informatycy robią testy na danych, na podstawie czego wiedzą, które elementy NER trzeba douczyć. Losują próbkę danych, którą wgrywają do systemu do anotacji. Lingwiści anotują próbkę, na podstawie której Informatycy douczają system. Od rodzaju kontraktu zależy, czy dane zostaną u nas, do dalszego douczania różnych systemów, czy Użytkownicy je usuną, czy udostępnią na licencji otwartej.

Przykład z anonimizacją

Użytkownik chcę udostępnić dane na mocy porozumienia z jednostką badawczą dane do douczania systemów jako swój wkład w rozwój Infrastruktury. Dane pochodzą z systemu czatowego i zawierają dane wrażliwe. Przed udostępnieniem muszą zostać zanonimizowane za pomocą narzędzia, które jest instalowane u Użytkownika. Następnie osoba delegowana czyści dane (drugie sprawdzenie). Dalsze zarządzanie odbywa się jak w przykładzie powyżej.

Zadania związane z zarządzaniem danymi oraz zasoby

Plan zarządzania danymi:

- jest elementem wniosku konkursowego, agendy badawczej lub studium wykonalności (w zależności od konkursu);
- w indywidualnych przypadkach pomiędzy stronami podpisywana jest umowa;
- współpracujące jednostki ustalają warunki zarządzania danymi;
- przed startem prac dane muszą zostać złożone w jednym z naszych systemów;
- plan zarządzania danymi jest aktualizowany przez Koordynatora współpracy niekomercyjnej (Asystenta Użytkowników) i Koordynatora współpracy komercyjnej.

Data Steward - osoba, która zna potrzeby i możliwości danej jednostki badawczej, daje swoje rekomendacje, bierze udział w negocjowaniu umów, współtworzy dokumenty wnioskowe. Data Steward jest pracownikiem Politechniki Wrocławskiej.

Zadania związane z zarządzaniem danymi oraz zasoby

Budżet na cele zarządzania danymi i zagwarantowanie przestrzegania zasad FAIR?

Zespół do zarządzania danymi:

- zespół DevOps
  - kompetencje: systemy repozytoryjno-obliczeniowe, ich obsługa i integracja danych i systemów informatycznych;
- koordynatorzy współpracy naukowej i komercyjnej
  - kompetencje: pozyskiwanie danych, umowy, licencje, planowanie dalszej współpracy w obrębie zespołu i pomiędzy zespołami;
- zespół Lingwistów
  - kompetencje: formaty i standardy danych, metadane, strukturyzacja danych, systemy do anotacji;
- zespół Informatyków - Programistów
  - kompetencje: formaty danych, normalizacja i anonimizacja danych wrażliwych, metadane, przetwarzanie danych tak, by były reużywalne.

Zadania związane z zarządzaniem danymi oraz zasoby

Budżet na cele zarządzania danymi i zagwarantowanie przestrzegania zasad FAIR?

Koszty:

- utrzymanie stanowisk osób, zajmujących się danymi;
- utrzymanie systemów (moc obliczeniowa, pojemność serwerowa);
- zapewnienie aktualnego oprogramowania i nowoczesnej infrastruktury sprzętowej;
- koszty współpracy z zespołami międzynarodowymi (np. w ramach sieci naukowych);
- koszty aktualizacji danych;
- obsługa prawna (u nas: po stronie Politechniki Wrocławskiej).

Zadania związane z zarządzaniem danymi oraz zasoby

Przykład

Słowosieć jest największym relacyjnym słownikiem dla języka polskiego. Jest ona połączona z analogicznym słownikiem dla języka angielskiego (Princeton WordNet) oraz siecią otwartych połączonych danych (Linked Open Data). Jest udostępniana na licencji Princeton WordNet. Sposoby opisu jednostek językowych w ramach sieci są standaryzowane, co zapewnia interoperacyjność - zgodność z formatami słowników dla innych języków. Słowosieć została ponadto zapisana w fomalizmie SKOS (uznanym przez W3C), co umożliwia połączenie ze światowymi połączonymi danymi - tezaurusami, słownikami i ontologiami.

Backup Słowosieci jest robiony codziennie automatycznie, zamknięte bazy opublikowanych wersji są przechowywane na serwerach Politechniki Wrocławskiej oraz dwóch innych uczelni.

Do obsługi Słowosieci, oprócz lingwistów, aktualizujących dane, potrzebni są informatycy do aktualizacji baz danych i aplikacji, korzystających z danych. Konieczna jest również znaczna moc obliczeniowa, która umożliwia utrzymanie i aktualizację grafu Linked Open Data.

Wdrażanie i raportowanie Planów Zarządzania Danymi

Wdrażanie i raportowanie Planu Zarządzania Danymi

- **DMP może zmieniać się w trakcie realizacji projektu (zalecane)**
- **Raport z wdrażania – należy opisać stan faktyczny na koniec realizacji projektu: planowano vs. zrealizowano.**

Jak przygotować się do opublikowania danych badawczych - 5 (6) kroków

1. Ustalenie zgody na udostępnienie danych (zgoda współtwórców)
2. Sprawdzenie regulacji w jednostce naukowej (ew. zgoda kierownika/dyrektora)
3. Selekcja materiału publikowanego: model otwarty lub ograniczony
4. Wybór odpowiedniego repozytorium
5. Deponowanie danych/metadanych (wraz z opisem, licencją)
6. (opcjonalnie) Przygotowanie publikacji wiodącej z opisem zasobu

Powyższe kroki należy wykonać “na sucho” podczas przygotowania DMP. W razie zmian podczas realizowania projektu, rozbieżności i decyzje powinny być dokumentowane i raportowane (w trakcie lub po zakończeniu projektu).

Kontakt CLARIN-PL / PWr:

agnieszka.dziob@pwr.edu.pl

jan.wieczorek@pwr.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.

NARODOWE CENTRUM

Ministerstwo Edukacji i Nauki

Narodowe Centrum Nauki

Zespół ds. Otwartej Nauki

otwarta.nauka@ncn.gov.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.

NARODOWE CENTRUM

Ministerstwo Edukacji i Nauki