Dziedzinowe repozytoria otwartych danych badawczych

Publiczna prezentacja założeń projektu
Warszawa, 8 grudnia 2016

Partnerzy projektu

- Instytut Filozofii i Socjologii, Polska Akademia Nauk
- Uniwersytet Adama Mickiewicza w Poznaniu: Wydział Chemii
- Uniwersytet Warszawski (lider projektu): Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego oraz Instytut Studiów Społecznych im. Profesora Roberta B. Zajonca

Plan prezentacji

1. Istota projektu
2. Diagnoza potrzeb użytkowników
3. Istota projektu - raz jeszcze
4. Cele i wskaźniki
5. Budżet projektu i czas realizacji
6. Harmonogram zamówień publicznych

Istota projektu

- **Przystosowanie** istniejącego **oprogramowania** do pełnienia funkcji repozytorium **ogólnego** zastosowania oraz jego wdrożenie.
- **Rozwinięcie funkcjonalności** dot. w.w. oprogramowania i przystosowanie go do pełnienia funkcji repozytoriów **dziedzinowych** (dane społeczne, dane krystalograficzne), a następnie ich wdrożenie.
- **Opracowanie schematów metadanych** dla repozytoriów.
- **Przygotowanie i udostępnienie danych i metadanych** dla uruchomionych trzech repozytoriów.
- Przeprowadzenie **warsztatów** dla członków grup docelowych.

Diagnoza potrzeb

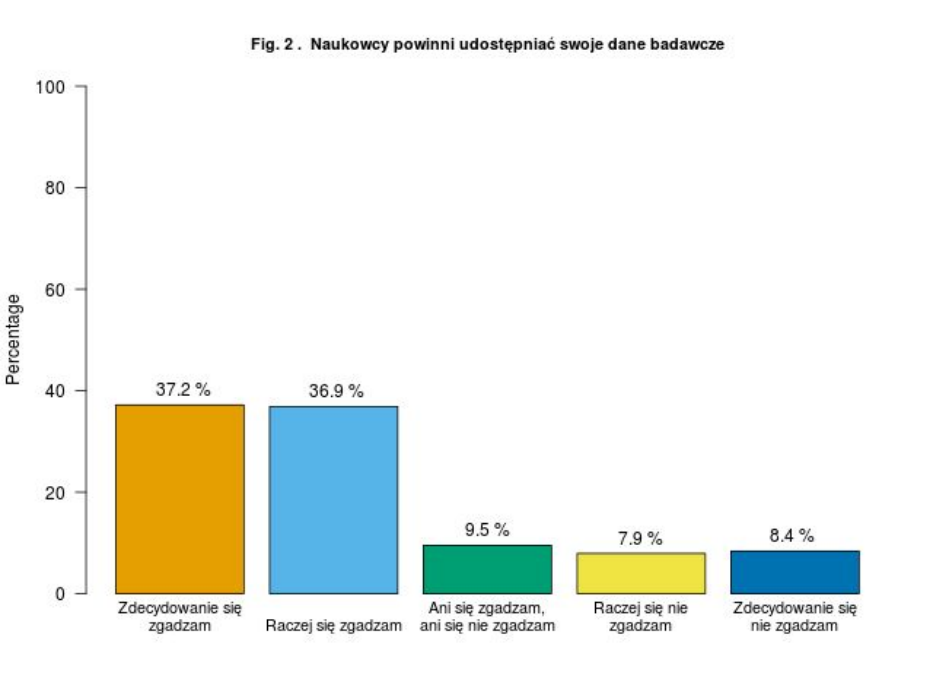
- Dane ilościowe - ankieta internetowa zrealizowana wśród naukowców, lipiec - sierpień 2015, zrealizowane dla potrzeb raportu “Towards open research data in Poland” (N=630).
- Dane jakościowe jakościowych - badanie metodą IDI (10 osób, II połowa 2015), naukowcy udostępniający swoje dane badawcze i korzystający z danych udostępnionych przez innych, zrealizowane dla potrzeb raportu “Towards open research data in Poland”.
- Informacji o użytkownikach oraz ruchu na stronach Archiwum Danych Społecznych.

Diagnoza potrzeb c.d.

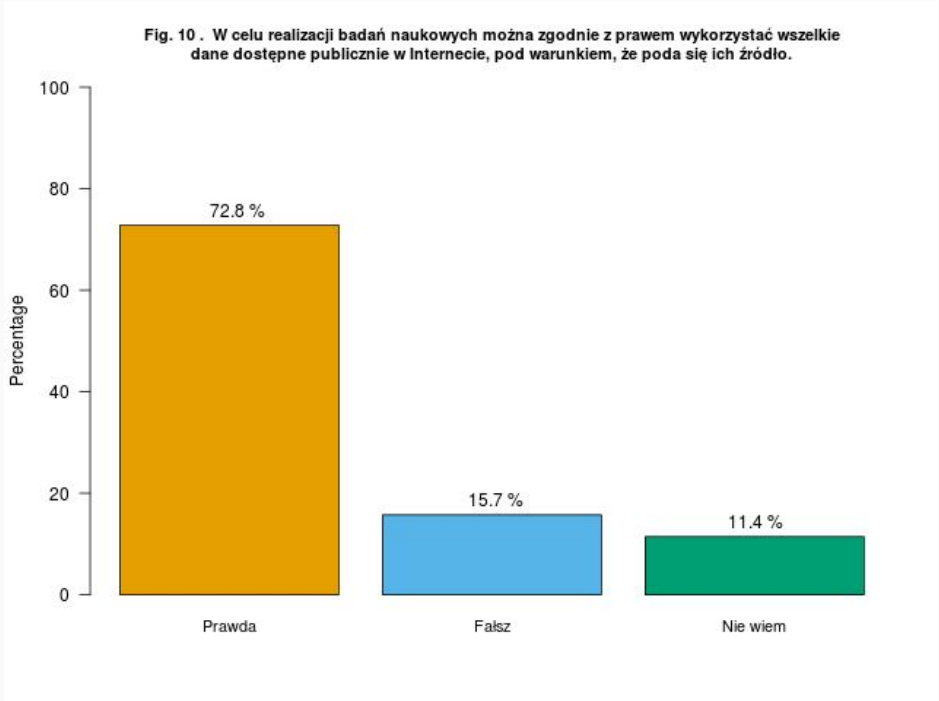
Badanie ilościowe:

- Przeważa ogólnie pozytywny stosunek do udostępniania danych badawczych.
- Większa ostrożność w przypadku bardziej szczegółowych pytań (wykorzystanie komercyjne, moment udostępnienia).
- Kłopot z prawnymi aspektami udostępniania.
- Dla naukowców ważny jest wpływ udostępnienia na rozwój kariery (w sensie pozytywnym i negatywnym).
- Wysoka deklarowana skłonność do wykorzystania danych udostępnionych przez innych (badania, dydaktyka, weryfikacja wyników).

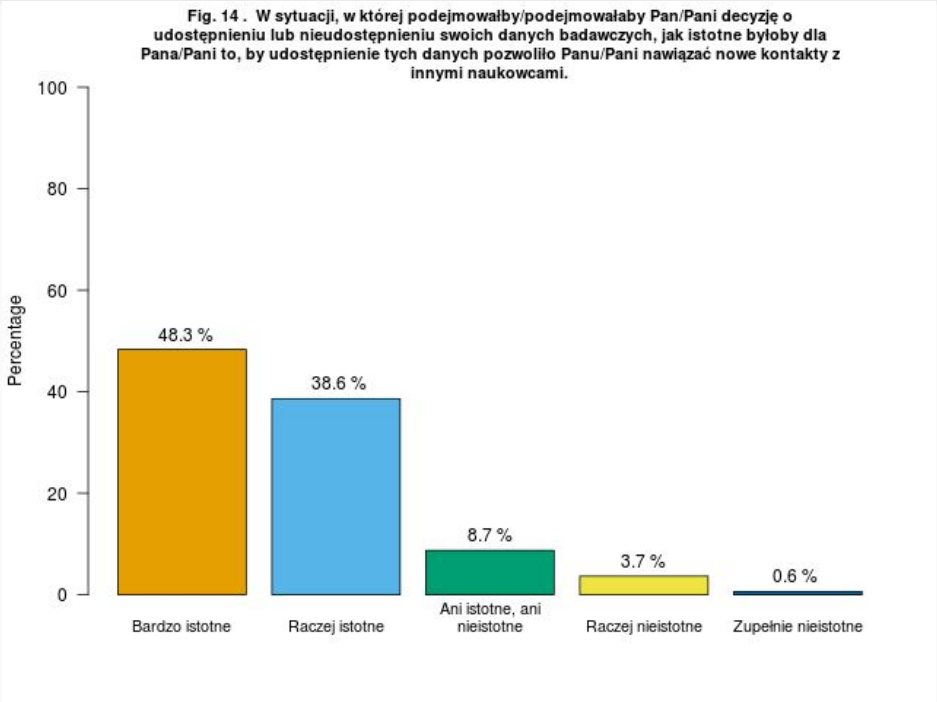
Diagnoza potrzeb c.d.



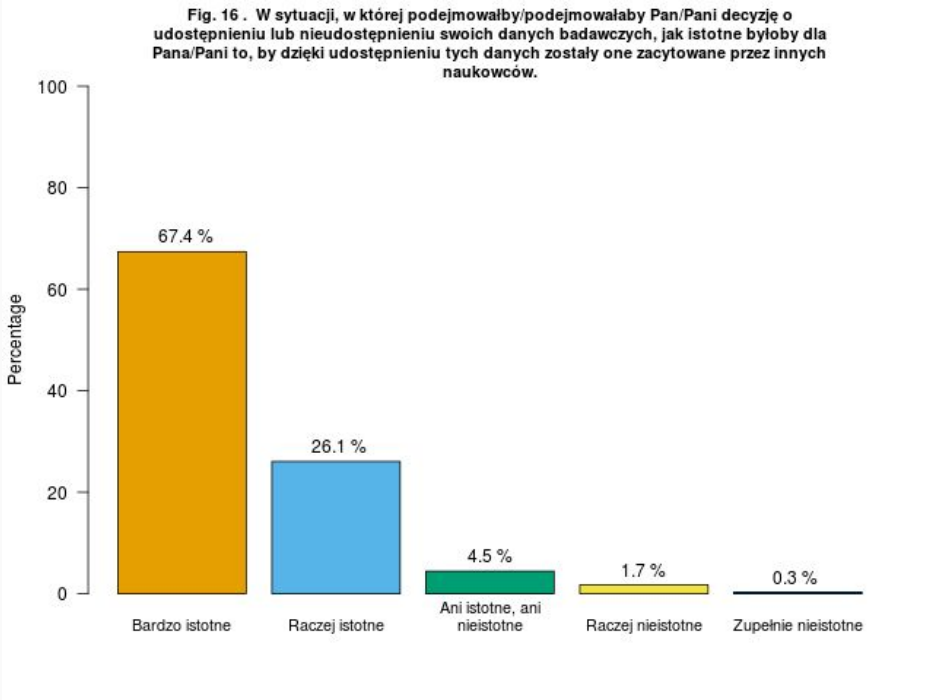
Diagnoza potrzeb c.d.



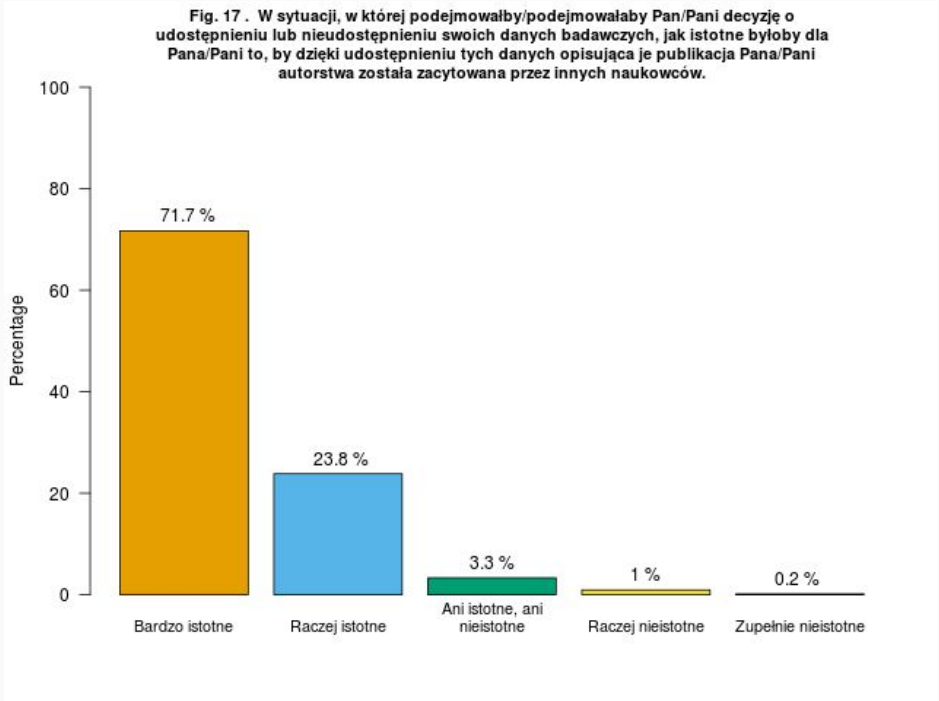
Diagnoza potrzeb c.d.



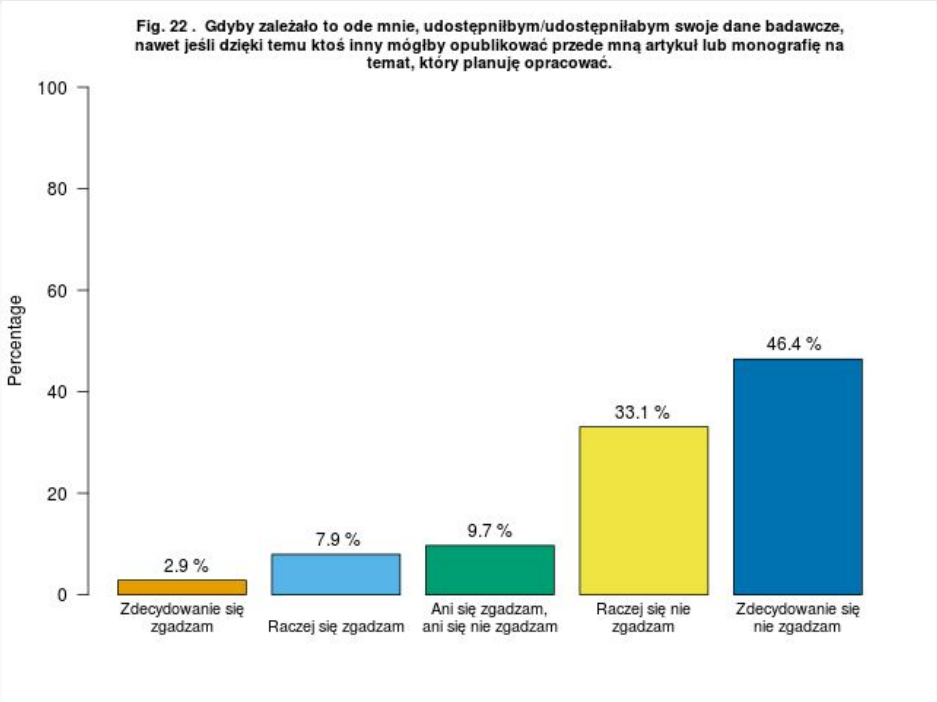
Diagnoza potrzeb c.d.



Diagnoza potrzeb c.d.



Diagnoza potrzeb c.d.



Diagnoza potrzeb c.d.

Analiza ruchu na stronie Archiwum Danych Społecznych:

- Przeważają użytkownicy ze świata akademickiego.
- Pracownicy naukowi, ale przede wszystkim studenci.
- Ale: również przedsiębiorcy, a także (w mniejszej skali) samorządy i instytucje samorządowe.

Diagnoza potrzeb c.d.

Badania jakościowe:

- “Wszystko płynie”: oczekiwania, dane, infrastruktura.
- Duża istotność nagród związanych z udostępnieniem danych.
- Słaba orientacja w prawnych aspektach udostępniania.
- Trudności na jakie napotykają naukowcy zależą od dziedziny (infrastruktura, ale i kwestie etyczne).

Diagnoza potrzeb c.d.

[...] *jeżeli na przykład bym je pozyskał od respondentów i one by pozwalały na łatwą identyfikację tych respondentów w jakiś sposób,* ***a jeszcze na dodatek zawierałyby informacje, właśnie tak ogólnie mówiąc, sensitive, takie wrażliwe,*** *takie, nie wiem, z poglądami politycznymi, seksualnymi, czy jakimikolwiek innymi, no to tutaj naprawdę bym musiał się bardzo mocno uważać, jeśli bym myślał o udostępnieniu takich danych.* [B1]

[...] ***To są naprawdę takie gigantyczne ilości danych****. I wielkość tych danych będzie rosła cały czas. Bo to każda nowa technologia, każdy nowy rodzaj [urządzenia], który wychodzi na rynek [...] ma z reguły lepsze parametry. Ale to się wiąże z tym, że te pliki rosną. I to rosną w sposób eksponencjalny bardziej.* [B2]

Zdiagnozowane potrzeby i grupy docelowe

- Grupy docelowe: **naukowcy, studenci, przedsiębiorcy**.
- Potrzeba **uzyskania dostępu** do danych naukowych oraz ich **wykorzystania**.
- Potrzeba stworzenia **infrastruktury**, która byłaby pojemna i była w stanie podążać za zmianami otoczenia.
- Potrzeba udostępniania danych w sposób pozwalający na **uzyskanie nagród** wynikających z udostępnienia danych i ich wykorzystania (dobrze opracowane dane i metadane, wymiana metadanych, DOI).
- Potrzeba uwzględnienia **specyfiki** poszczególnych dziedzin.
- Potrzeba **podniesienia poziomu wiedzy** związanej z udostępnianiem i wykorzystywaniem danych (wsparcie przez działania miękkie).

Istota projektu

- **Przystosowanie** istniejącego **oprogramowania** do pełnienia funkcji repozytorium **ogólnego** zastosowania oraz jego wdrożenie.
- **Rozwinięcie funkcjonalności** dot. w.w. oprogramowania i przystosowanie go do pełnienia funkcji repozytoriów **dziedzinowych** (dane społeczne, dane krystalograficzne), a następnie ich wdrożenie.
- **Opracowanie schematów metadanych** dla repozytoriów.
- **Przygotowanie i udostępnienie danych i metadanych** do uruchomionych trzech repozytoriów.
- Przeprowadzenie **warsztatów** dla członków grup docelowych.

Cele projektu

Cel bezpośredni 1. **Poprawa cyfrowej dostępności** zasobów nauki za pośrednictwem repozytoriów otwartych danych naukowych

Cel pośredni 1. Zwiększenie dostępności danych naukowych poprzez **opracowanie** w oparciu o istniejące rozwiązania **oprogramowania pełniącego rolę repozytorium otwartych danych** naukowych ogólnego zastosowania oraz **repozytoriów dziedzinowych**.

Cel pośredni 2. **Wdrożenie** trzech instancji opracowanego oprogramowania.

Cel pośredni 3. **Zamieszczenie i udostępnienie** w bazach danych wdrożonych repozytoriów danych naukowych wraz z metadanymi.

Cele projektu - c.d.

Cel bezpośredni 2. **Poprawa jakości** udostępnianych danych naukowych oraz ich metadanych

Cel pośredni 4. **Opracowanie danych** naukowych w sposób umożliwiający ich udostępnienie i ponowne wykorzystanie.

Cel pośredni 5. Poprawa jakości danych poprzez **opracowanie i wdrożenie schematów metadanych** dla repozytoriów.

Cel pośredni 6. Poprawa jakości danych poprzez **opracowanie metadanych** dla udostępnianych zbiorów zgodnie z wypracowanymi standardami.

Cele projektu - c.d.

Cel bezpośredni 3. **Lepsze wykorzystanie** udostępnionych danych naukowych

Cel pośredni 7. Zwiększenie wykorzystania udostępnionych w projekcie danych dzięki objęciu ich **wolnymi licencjami**.

Cel pośredni 8. **Lepsza integracja** udostępnionych danych z usługami zewnętrznymi dzięki wymianie metadanych z serwisami zewnętrznymi.

Cel pośredni 9. **Podniesienie wiedzy i świadomości** w zakresie udostępniania i wykorzystania danych naukowych wśród kluczowych grup interesariuszy.

Wskaźniki

Cel bezpośredni 1. **Poprawa cyfrowej dostępności** zasobów nauki za pośrednictwem repozytoriów otwartych danych naukowych	1. Liczba podmiotów, które udostępniły on-line informacje sektora publicznego: 3. 2. Liczba pobrań udostępnionych plików: 21 000. 3. Rozmiar danych udostępnionych online: 3-5 TB. 4. Liczba udostępnionych zbiorów danych: 600. 5. Liczba opracowanych aplikacji: 3. 6. Liczba wdrożonych instancji repozytoriów: 3.
Cel bezpośredni 2. **Poprawa jakości** udostępnianych danych naukowych oraz ich metadanych	1. Liczba opracowanych schematów metadanych: 3. 2. Liczba repozytoriów, w których wdrożono opracowane schematy metadanych: 3. 3. Liczba opracowanych zbiorów danych: 600. 4. Liczba opracowanych zestawów metadanych: 600.
Cel bezpośredni 3. **Lepsze wykorzystanie** udostępnionych danych naukowych	1. Liczba zbiorów udostępnionych na wolnych licencjach: 200. 2. Liczba utworzonych API: 3. 3. Liczba przeprowadzonych szkoleń: 12.

Zasoby objęte projektem i ich znaczenie - dane krystalograficzne

Ok. **200** zbiorów surowych danych dyfrakcyjnych, służące analizie struktur białek. Badania tego rodzaju stanowią podstawę m.in. dla nowych terapii lekowych.

Większa dostępność danych tego rodzaju umożliwi **weryfikację** uzyskanych wcześniej wyników oraz ułatwi tworzenie nowych **algorytmów** służących do analizy tego rodzaju danych.

Zasoby objęte projektem i ich znaczenie - dane społeczne

Ok. **400** zbiorów danych społecznych, stanowiących zarówno rezultat badań ilościowych, jak i jakościowych.

Dane wykorzystywane są zarówno do celów **dydaktycznych**, jak i do celów **badawczych** (analiza danych zastanych), stanowiąc podstawę dla nowych analiz i publikacji naukowych. Dane mogą zostać wykorzystane również przez **podmioty pozaakademickie** (przedsiębiorstwa, dziennikarzy, NGO, instytucje rządowe i samorządowe).

Zasoby objęte projektem i ich znaczenie - pozostałe dane

Ok. **20** zróżnicowanych zbiorów danych mających zasilić repozytorium ogólnego zastosowania.

Ze względu na ich zróżnicowany charakter, najważniejsze korzyści związane z ich udostępnieniem dotyczyć będą przede środowiska akademickiego: możliwości **weryfikacji** uzyskanych wyników, **powtórnego wykorzystania** danych do celów badawczych oraz wykorzystania ich w celach **dydaktycznych**.

Najważniejsze korzyści

- Akademickie:
  - Możliwość weryfikacji wyników, badania oryginalne.
  - Ograniczenie kosztów badań - brak konieczności powtórnej realizacji badań.
  - Ułatwienie współpracy z czasopismami i instytucjami finansującymi wymagającymi deponowania danych.
  - Poprawa stanu wiedzy na temat udostępniania i wykorzystywania danych (szkolenia).
- Ekonomiczne:
  - Możliwość weryfikacji danych krystalograficznych na wczesnym etapie umożliwi ograniczenie kosztów badań nad nowymi lekami.
- Społeczne:
  - Możliwość wykorzystania danych do celów popularyzatorskich i edukacyjnych (zwłaszcza repozytorium danych społecznych).
  - Podniesienie jakości kształcenia akademickiego.

Czas realizacji i koszt projektu

Czas realizacji projektu: 36 miesięcy (lipiec 2017 - czerwiec 2020)

Koszt projektu: 4 998 800 PLN

Harmonogram zamówień publicznych

	Forma	Termin realizacji	Wartość
Zakupy sprzętu komputerowego	Przetarg europejski	Miesiące 2-5	123 000 PLN
Projekty graficzne	Tryb negocjacji	Miesiące 5-8	74 000 PLN
Catering (warsztaty)	12 zapytań ofertowych przed każdym warsztatem	Miesiące 14-16, 22-24, 28-31, 34-35.	20 000 PLN

Dziękujemy za uwagę

w.fenrich@icm.edu.pl