PLATFORMA OTWARTEJ NAUKI

Otwarte dane badawcze w humanistyce

Marta Hoffman-Sommer
Michał Starczewski

17 marca 2016 r.

UNIWERSYTET WARSZAWSKI
Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego
www.icm.edu.pl

PLATFORMA OTWARTEJ NAUKI

Plan na dziś

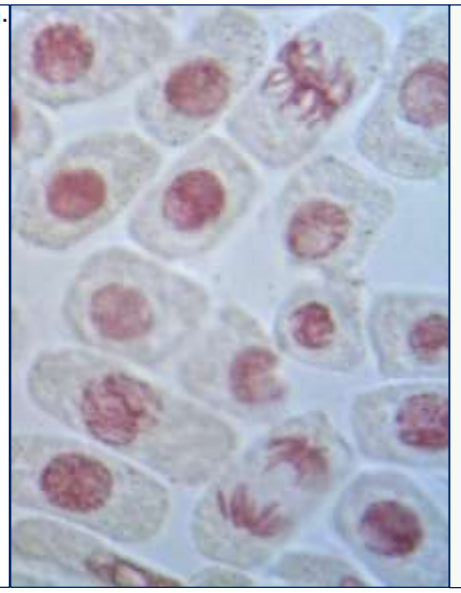
- Czym są otwarte dane badawcze?
- Zarządzanie danymi badawczymi w 5 krokach
- Plan ZDB
- Gdzie szukać otwartych danych?
- Kwestie prawne

Czym są dane badawcze w humanistyce?

Rezultaty badań naukowych

Artykuły i książki

Channel	Raw Int.	Intensity		Avg.
11481	61,73	69	186	60
42142	181,65	447	232	73
37539	151,37	403	248	67
26707	127,18	302	210	69
33831	145,82	329	232	71
30312	135,32	310	224	73
20118	83,82	125	240	71
16894	83,22	140	203	71
16143	82,36	115	196	71
19950	95	159	210	73
24331	98,11	174	248	75
21530	106,06	222	203	71
11831	67,99	77	174	67
46601	194,17	428	240	79
52345	180,5	468	290	81
43917	177,08	428	248	77
43813	208,63	478	210	83
39835	177,83	422	224	81
20207	103,1	170	196	77
17899	91,32	136	196	75
15462	88,86	136	174	73
18585	94,82	155	196	74
21416	109,27	197	196	79
26097	112,49	212	232	77
11463	63,68	73	180	65
36909	144,18	277	256	77
40585	145,47	293	279	75
32514	140,15	256	232	79
38101	127	283	300	77
29338	104,78	203	280	73
26193	93,88	144	279	77



```
<TEI version="5.0" xmlns="http:/
 <teiHeader>
  <fileDesc>
   <titleStmt>
    <title>TEI中文指引</title>
   </titleStmt>
   <publicationStmt>
    <p>將與TEI 中文在地化計劃等文件一
   </publicationStmt>
   <sourceDesc>
    <p>譯自TEI P5 英文指引</p>
   </sourceDesc>
  </fileDesc>
 </teiHeader>
 <text>
  <body>
   <p>這是TEI P5的中文指引...</p>
  </body>
 </text>
</TEI>
```

Dane badawcze

Dane badawcze:

„…zarejestrowane materiały o charakterze faktograficznym powszechnie uznawane przez społeczność naukową za niezbędne do oceny wyników badań naukowych.”

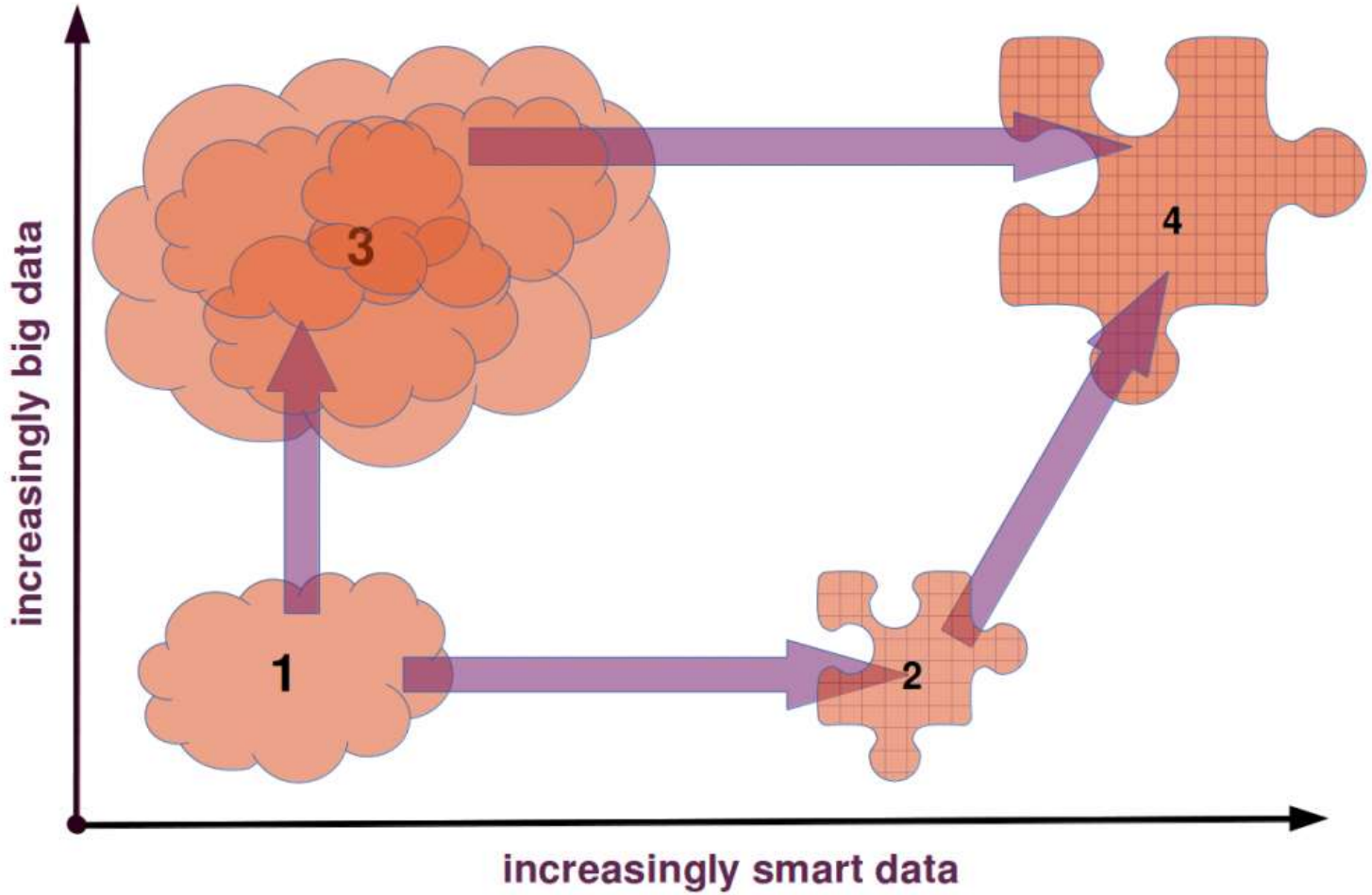
Co zaliczamy do danych badawczych?

- Dokumenty tekstowe
- Notatki
- Kwestionariusze, ankiety, wyniki badań ankietowych
- Nagrania audio, wideo
- Fotografie
- Oprogramowanie
- Korpusy językowe
- Archiwa mediów społecznościowych (Twitter)
- Dane liczbowe
- Obiekty
- …

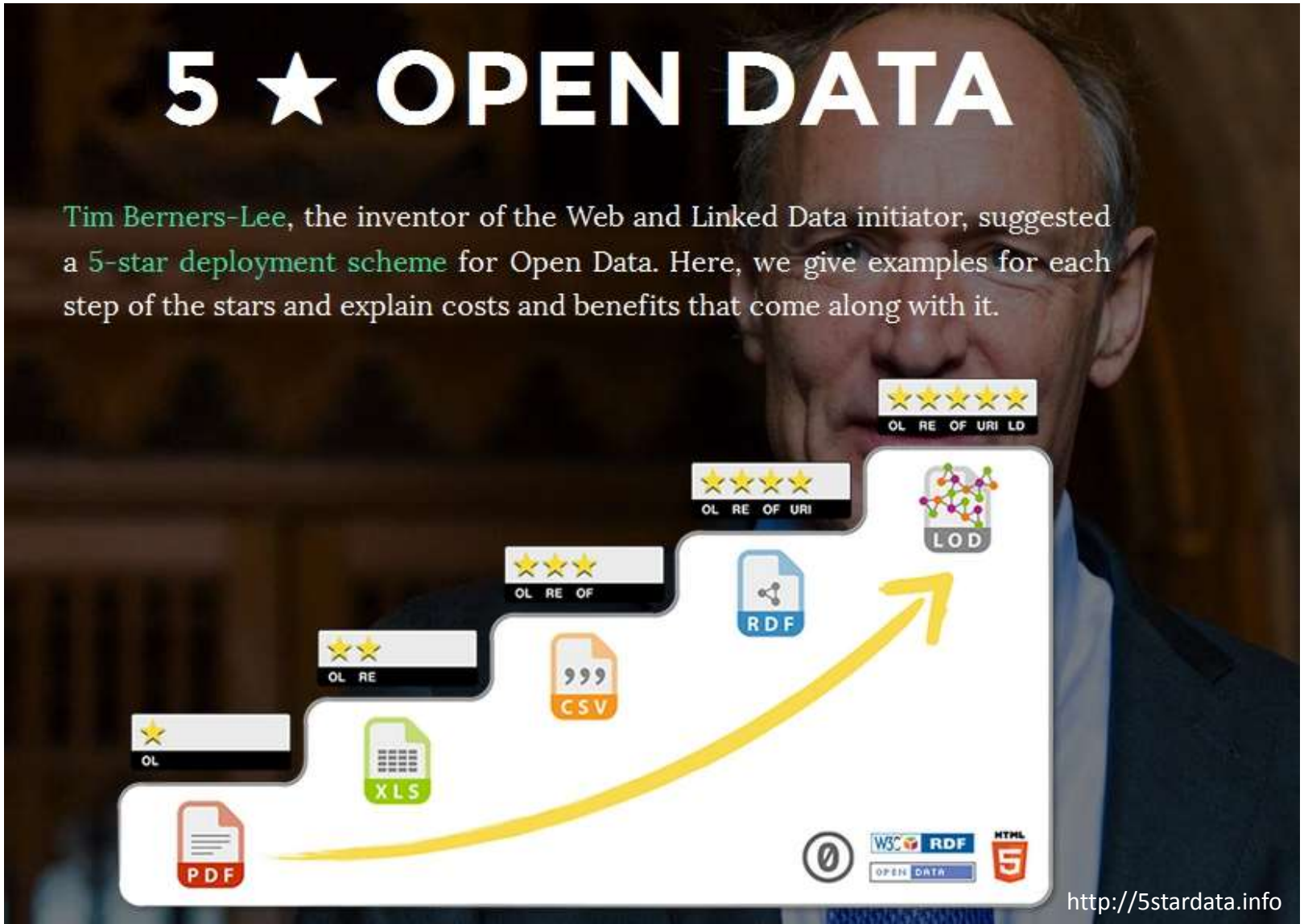
http://europeana.eu/portal/record/2026117/Partage_Plus_ProvidedCHO_Museu_Nacional_d_Art_de_Catalunya_000416_C.html

Big data – smart data

- Christof Schöch http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/



Otwarte dane badawcze

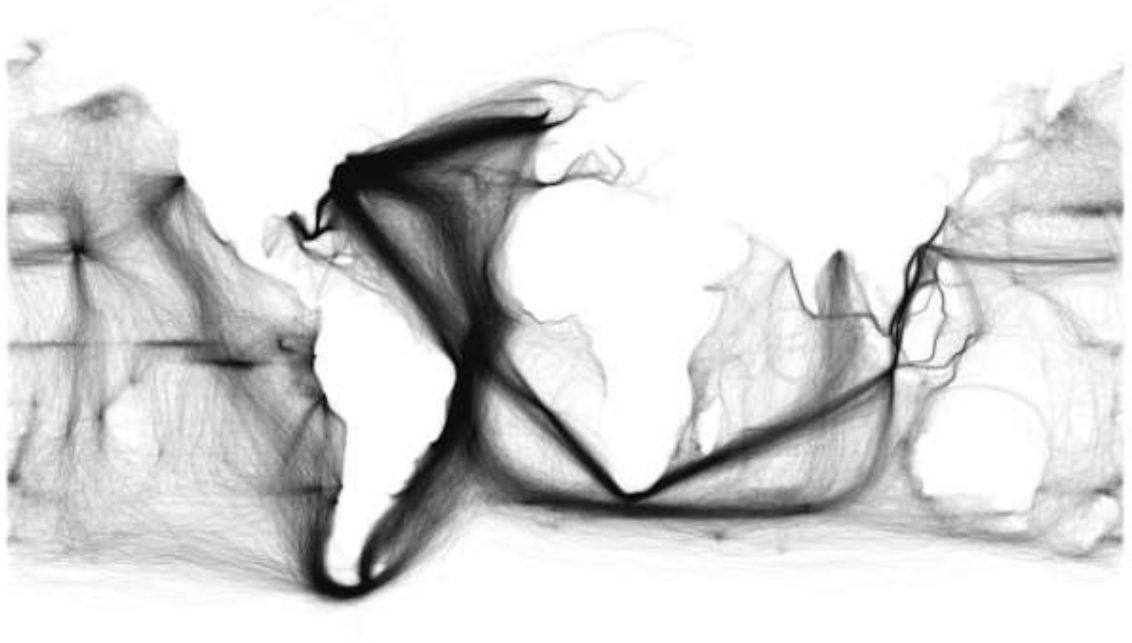


PLATFORMA
OTWARTEJ
NAUKI

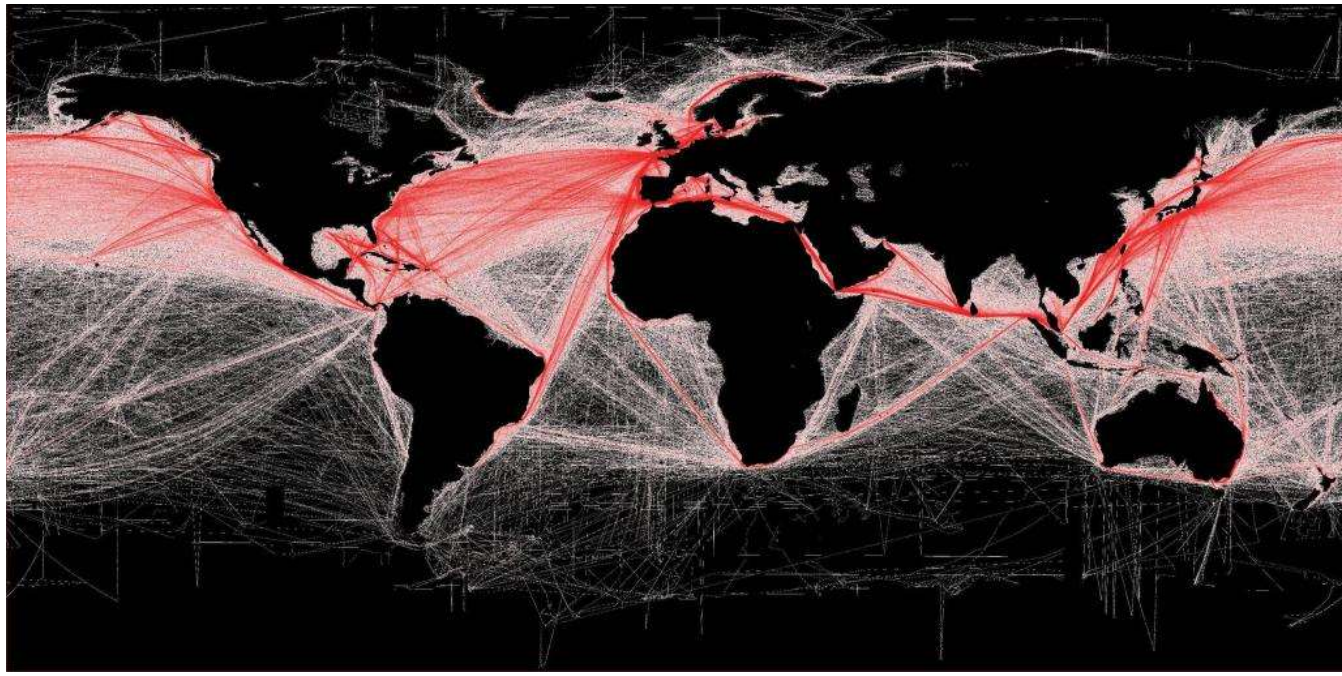
OPEN DATA FOR BETTER RESEARCH

OTWARTE DANE, LEPSZA NAUKA

Po co otwierać dane badawcze?



All voyages from the ICOADS US Maury collection. Ships tracks in black, plotted on a white background, show the outlines of the continents and the predominant tracks on the trade winds. Original source here.



http://www.zmescience.com/other/feature-post/shipping-wind-boat/

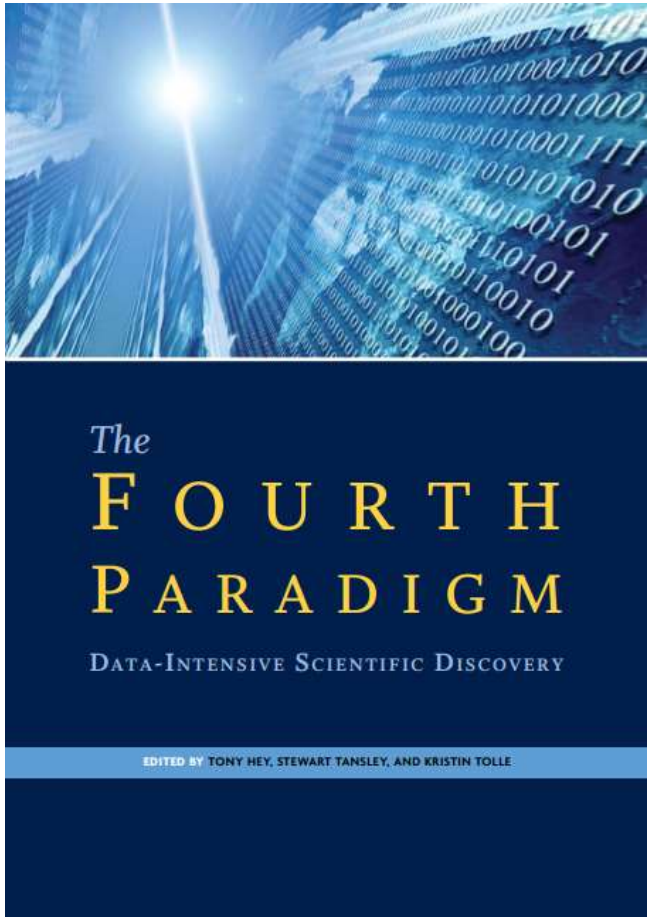
Po co udostępniać dane?

- Weryfikacja wyników
- Kolejne badania
- Nowe kontakty naukowe

Zmiany w sposobie uprawiania nauki

Cztery paradygmaty w nauce (Jim Gray, 2007):

Empiryczny – opis zjawisk naturalnych
(ostatnie tysiaclecia)
Teoretyczny – budowa modeli i uogólnień
(ostatnie stulecia)
Obliczeniowy – symulacje złożonych zjawisk
(ostatnie dekady)
Eksploracja danych – badania „data-intensive”, w tym analiza maszynowa (text mining, data mining)
(ostatnie lata)



Sytuacja w Polsce

Dokument MNiSW (październik 2015):

Kierunki rozwoju otwartego dostępu do publikacji i wyników badań naukowych w Polsce

„…zaleca, aby krajowe podmioty finansujące badania naukowe ze środków publicznych (…) stosowały i upowszechniały zasady, zgodnie z którymi publikacje i **dane badawcze** powstające w wyniku finansowanych lub współfinansowanych przez nie badań **znajdą się w otwartym dostępie**.”

Wymagania Komisji Europejskiej w programie Horyzont 2020

Pilotaż Otwartych Danych w H2020

Pilotaż Otwartych Danych Badawczych:

„Od finansowanych projektów wchodzących w zakres objęty Pilotażem Otwartych Danych Badawczych jest wymagane korzystanie ze szczegółowego planu zarządzania danymi, odnoszącego się do poszczególnych zbiorów danych.”

„Pilotaż Otwartych Danych obejmuje dwa rodzaje danych:

1) **dane** (…) **niezbędne do weryfikacji wyników** prezentowanych w publikacjach naukowych należy udostępniać tak szybko, jak to możliwe;
2) **inne dane** (…) wymienione w planie zarządzania danymi należy udostępniać zgodnie z ustalonymi w planie terminami.

(…) Projekty objęte pilotażem są zobowiązane do deponowania opisanych powyżej danych badawczych, najlepiej w repozytoriach danych badawczych.”

Zarządzanie danymi badawczymi

...aktywne podejście do danych badawczych na wszystkich etapach ich cyklu życiowego.

Co uwzględnić?

1. Pozyskiwanie danych, dobór formatów plików, nazewnictwo plików, metadane, dokumentacja
2. Krótko- i długoterminowe przechowywanie danych: selekcja danych, bezpieczna archiwizacja
3. Zasady dostępu do danych, możliwości ich ponownego wykorzystania
4. Prawne i etyczne aspekty rozporządzania zbiorem danych
5. Zasoby potrzebne do zarządzania danymi (np. finansowe, kompetencje)

Jakie korzyści daje świadome ZDB?

1. ułatwienie dla własnych przyszłych badań
2. możliwość udostępnienia innym zainteresowanym
3. poprawa jakości uprawianej na świecie nauki
4. więcej współpracy w nauce
5. szybszy postęp w badaniach
6. oszczędność środków finansowych w nauce

Kroki do wykonania

1. Identyfikacja danych w projekcie
2. Bieżące zarządzanie danymi
3. Selekcja danych
4. Przygotowanie danych do archiwizacji
5. Deponowanie danych

1. Zidentyfikowanie danych

Skąd się biorą dane w naszym projekcie?

Jak często pojawiają się nowe dane?

Jak dużo danych powstaje w projekcie?

W jakich formatach są gromadzone dane?

Na podstawie: Workbook for Writing a Data Management Plan,
http://www.dcc.ac.uk/training/digital-curation-101/dmp-workshop-uct

2. Zarządzanie w trakcie projektu

Jakie stosujemy formaty oraz nazewnictwo plików i folderów?

Jakie dodatkowe informacje mogą być potrzebne do korzystania

z tworzonych danych (dokumentacja)?

Gdzie przechowujemy nasze dane na bieżąco?

W jaki sposób je zabezpieczamy (backupy, regulacja dostępu)?

Na podstawie: Workbook for Writing a Data Management Plan,
http://www.dcc.ac.uk/training/digital-curation-101/dmp-workshop-uct

3. Selekcja danych do archiwizacji: co przechowywać, co wyrzucać

Jakie dane chcemy przechowywać po zakończeniu projektu?

Gdzie zdeponujemy dane do przechowywania długoterminowego?

Jak długo będziemy je przechowywać?

Kto będzie miał do nich dostęp i na jakich zasadach?

Na podstawie: Workbook for Writing a Data Management Plan, http://www.dcc.ac.uk/training/digital-curation-101/dmp-workshop-uct

Wskazówki do selekcji danych

1. ***Wymagania prawne*** zobowiązujące nas do archiwizacji danych.
2. ***Wartość naukowa lub historyczna***: tu musimy rozważyć potencjalne zainteresowanie w przyszłości.
3. ***Wyjątkowość***: czy nasze dane duplikują się z innymi istniejącymi zbiorami danych?
4. ***Możliwość replikacji***: czy można takie dane ponownie zebrać? (wysokie koszty, jednorazowe wydarzenie)
5. ***Możliwość wykorzystania***: jakość i używalność danych (czy formaty są od strony technicznej dobrze dobrane? czy kwestie praw własności intelektualnej są wyjaśnione?)
6. ***Kwestie ekonomiczne***: koszty zarządzania danymi i przechowywania ich są uzasadnione w świetle potencjalnych przyszłych zastosowań.
7. ***Pełna dokumentacja***: dokumentacja jest poprawna i kompletna.

Na podstawie: Whyte, A. & Wilson, A. (2010). "How to Appraise and Select Research Data for Curation". DCC How-to Guides. Edinburgh: Digital Curation Centre.
Available online: http://www.dcc.ac.uk/resources/how-guides/appraise-select-data

Dane, których nie zamierzamy przechowywać

Dokumentować:

Co, dlaczego i kiedy zostało wyrzucone

Jakość danych

Potrzebuję tych danych natychmiast!!!
Nieważne, że nie są wyczyszczone – sam sobie poradzę!

Zmarnowałem już kawał życia czyszcząc i porządkując kiepskie dane od innych. Dopóki nie będą wyczyszczone i udokumentowane, nie interesują mnie. A w ogóle to mam teraz inne sprawy na głowie…

Slajd: Kevin Ashley, DCC, CC-BY

Dane badawcze nigdy nie są idealne.

Przechowujmy takie dane, które są **wystarczająco dobre**.

Ważne:
Opisujmy i dokumentujmy wszystkie wady i braki naszych danych!

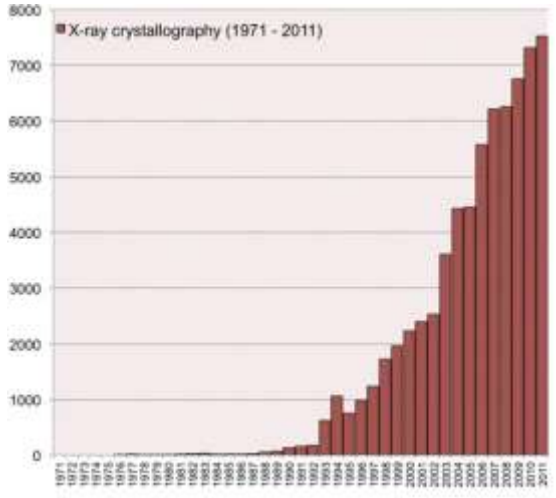
Gdzie przechowywać dane?

Cyfrowe repozytoria danych:

- specjalistyczne
- instytucjonalne
- szeroko zakrojone tematycznie
- ogólne

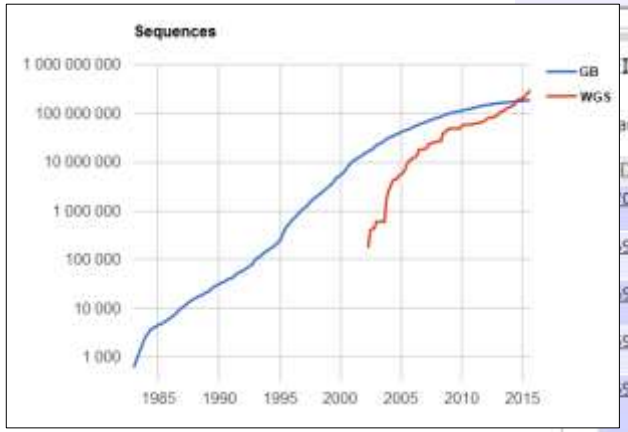
Repozytoria specjalistyczne

Protein Data Bank – od roku 1971



Berman, Kleywegt, Nakamura, Markley (2012)
http://dx.doi.org/10.1016/j.str.2012.01.010

GenBank – od roku 1982



http://www.ncbi.nlm.nih.gov/genbank/statistics

Oxford Text Archive – od roku 1976

Repozytoria tematyczne

Repozytorium danych biologicznych, dostępne dla wszystkich

Repozytorium danych z nauk społecznych i humanistycznych

Repozytoria instytucjonalne

Repozytorium uczelniane

Repozytorium tematyczne prowadzone przez brytyjską instytucję finansującą badania: Natural Environment Research Council

Repozytoria ogólne

Krajowe repozytorium danych: Holandia

Repozytorium ogólnodostępne
(publikacje + dane)

Krajowe repozytorium danych: Polska

Repozytorium ogólnodostępne
(publikacje + dane)

re3data.org – wyszukiwarka repozytoriów

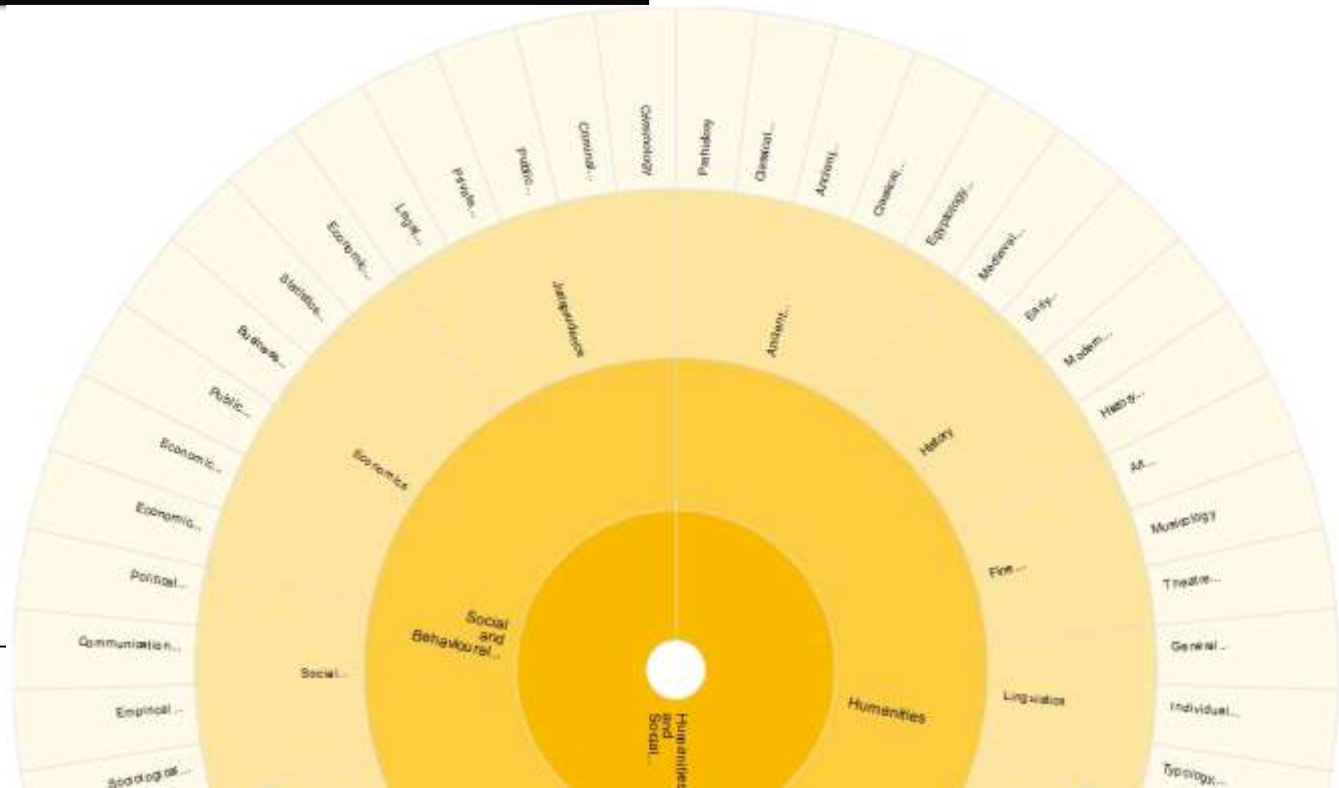
re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

Home | Search | **Browse** | Suggest | FAQ | About | Schema | API | Contact | Legal notice / Impressum

Browse

re3data.org supports the following browsing options:

- Browse by subject
- Browse by content type
- Browse by country



Czy wszystkie dane powinny być otwarte? Nie.

Dane osobowe

Bezpieczeństwo narodowe

Ochrona gatunków zagrożonych, stanowisk archeologicznych, etc.

Komercjalizacja wyników badań

Ale **informacja o istnieniu** danych zawsze powinna być publicznie dostępna:

- Inni mogą się dowiedzieć o danych i negocjować z nami dostęp
- Pozwala to uniknąć duplikacji badań

Slajd na podstawie: Kevin Ashley, DCC, CC-BY

4. Przygotowanie danych

Przygotowanie plików (ew. anonimizacja danych)

Metadane

Dokumentacja

Dobór formatów plików do archiwizacji (1)

Preferowane są formaty:

- Bez kompresji
- Nie wymagające komercyjnego oprogramowania
- Otwarte, z dostępną dokumentacją
- Wykorzystujące standardowe kodowanie (ASCII, Unicode)

Type	Recommended	Non-preferred
Tabular data	CSV, TSV, SPSS portable	Excel
Text	Plain text, HTML, RTF PDF/A only if layout matters	Word
Media	Container: MP4, Ogg Codec: Theora, Dirac, FLAC	Quicktime H264
Images	TIFF, JPEG2000, PNG	GIF, JPG
Structured data	XML, RDF	RDBMS

Na podstawie: UK Data Archive (nauki społeczne i humanistyczne)
http://www.data-archive.ac.uk/create-manage/format/formats

Dobór formatów plików do archiwizacji (2)

- Na bieżąco pracujemy w formatach, które nam najbardziej pasują – natomiast przed archiwizacją przenosimy pliki do standardowych, otwartych formatów.

- Niektóre repozytoria zachęcają do deponowania dwóch wersji tych samych danych:

  (1) w formacie przeznaczonym do długotrwałej archiwizacji,

  (2) w formacie najpowszechniej wykorzystywanym w danym środowisku.

Dokumentacja i metadane

Metadane: podstawowe informacje stanowiące opis całego zbioru danych (autor, tytuł, data powstania, nadana licencja, etc.)

Dokumentacja: informacje metodologiczne, kontekst powstania, dodatkowe pliki potrzebne do skorzystania z danych (skrypty), wykorzystane standardowe słowniki, etc.

Metadata standards: na stronach

Digital Curation Centre

www.dcc.ac.uk/resources/metadata-standards

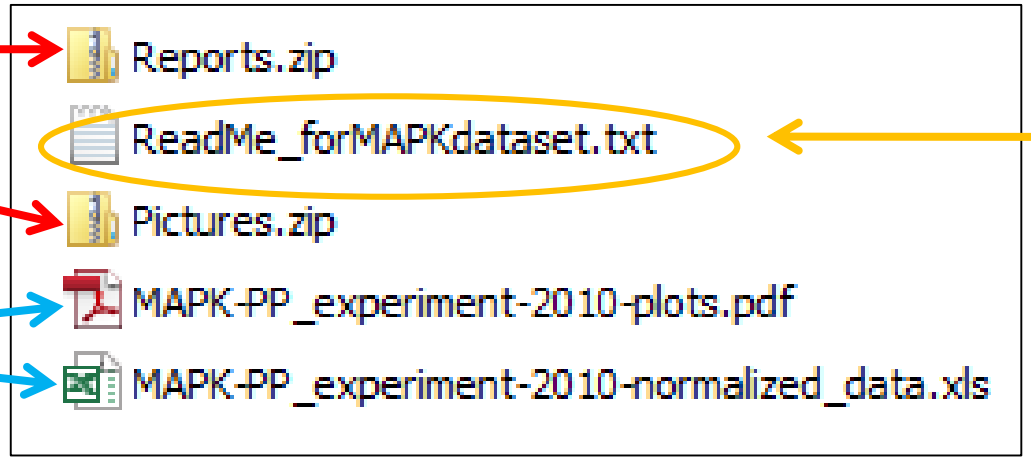
Slajd: Kevin Ashley, DCC, CC-BY

5. Deponowanie danych

Surowe dane: .txt

Dane przetworzone: .jpeg

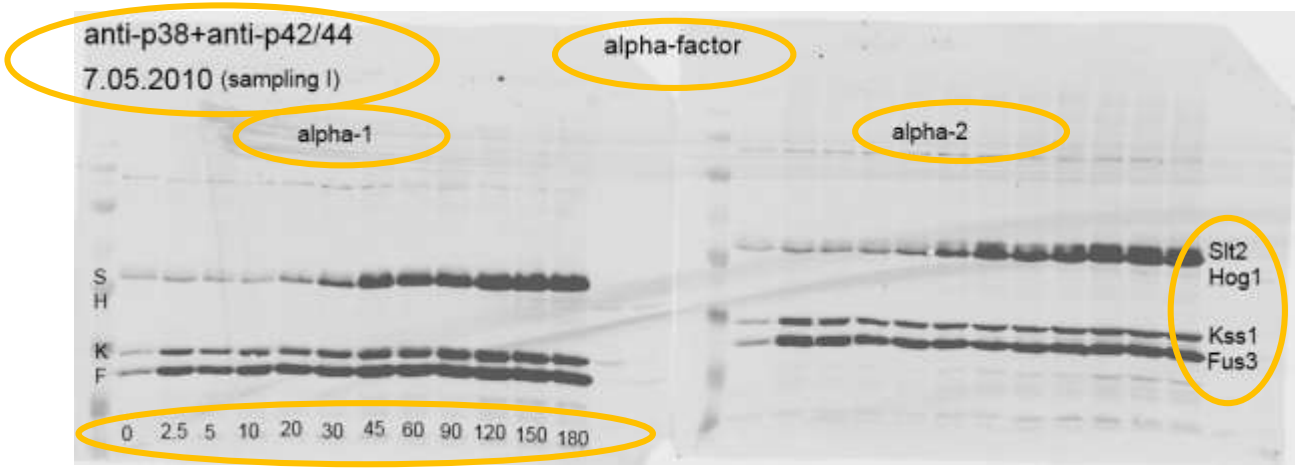
Analizy danych:
.xls, .pdf



Dokumentacja w osobnym pliku

Reports: oryginalne pliki z urządzenia pomiarowego, opisy w osobnym pliku

Pictures: tylko do weryfikacji wzrokowej, nie do analizy, opisy częściowo wpisane w pliki



Id	Name	I.I.(K Counts)	Shape Area	Channel	Raw Int. Intensity
0	totHog1-0-10_01	23,33	16,52	700	1208319
1	totHog1-0-10_02	24,5	16,52	700	1232882
2	totHog1-0-10_03	28,09	14,68	700	1319493
3	totHog1-0-10_04	19,15	12,9	700	939515
4	totHog1-0-10_05	12,43	10,81	700	667792
5	totHog1-0-10_06	29,86	13,76	700	1375125
6	totHog1-0-10_07	22,7	13,76	700	1118924
7	totHog1-0-10_08	31,32	14,68	700	1442892
8	totHog1-0-10_09	24,49	12,9	700	1158135
9	totHog1-0-10_10	20,58	13,33	700	1011508

RepOD - serwis dla polskiej społeczności akademickiej

repod.pon.edu.pl

Dane:

(1) badawcze

(2) otwarte

→ ze wszystkich dziedzin nauki

→ wszystkie formaty plików

Plan Zarządzania Danymi

Co to jest Plan Zarządzania Danymi (DMP)?

Krótki plan opisujący:

- Jakie dane zostaną wytworzone i w jaki sposób
- Jak te dane będą zarządzane (przechowywanie, zabezpieczanie, dostęp…)
- W jaki sposób będą archiwizowane i udostępniane innym

Co powinien zawierać plan DMP?

1. Jakie dane zostaną wytworzone lub zebrane?

   (co będą zawierać? jakie będą formaty plików? jak dużo będzie danych?)

4. Jak zostaną uporządkowane i opisane?

   (metadane, dokumentacja)

2. Kwestie etyczne i prawne

   (kwestie związane z ochroną prywatności, dane niejawne, etc.)

4. W jaki sposób dane zostaną udostępnione?

   (jak, kiedy, komu)

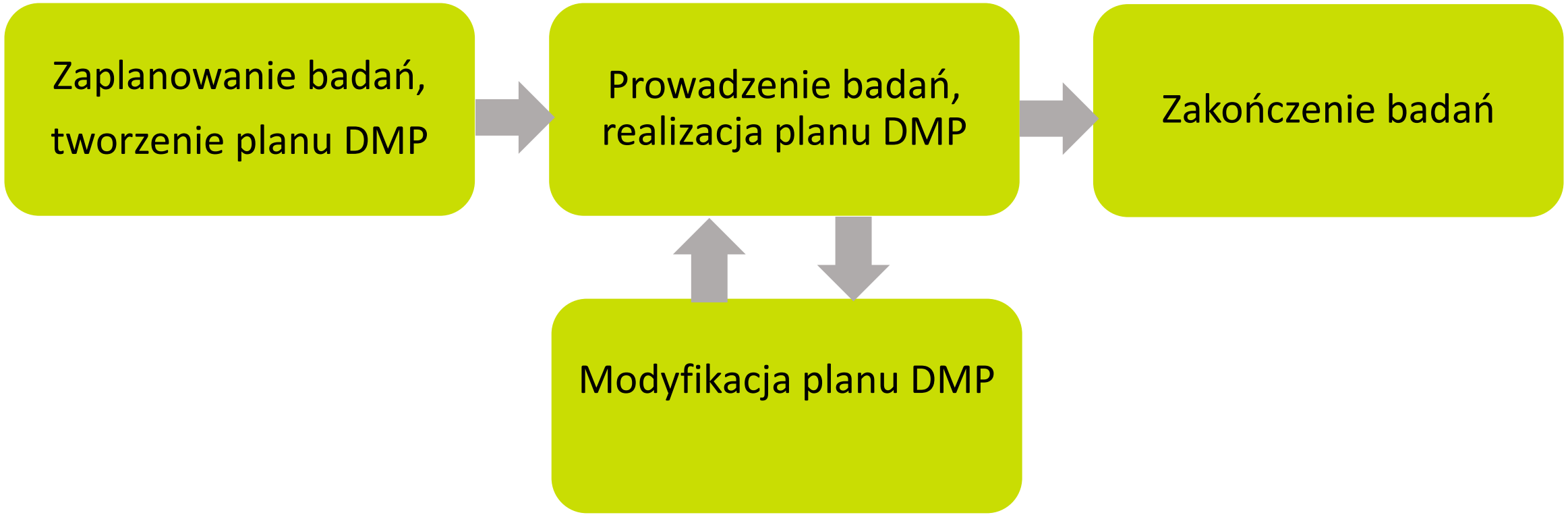
5. Które dane będą przechowywane długoterminowo? Gdzie, jak długo?

Slajd na podstawie materiałów DCC:
www.dcc.ac.uk/resources/data-management-plans/checklist

Jak napisać dobry plan DMP

- Plan powinien być krótki i prosty, ale konkretny
- Szukajmy wsparcia – konsultujmy się i współpracujmy
- Oprzyjmy nasz plan na dostępnych nam umiejętnościach i dostępnym wsparciu
- Plan powinien być realistyczny
- **Pamiętajmy**: plan może się zmieniać, ewoluować

Zarządzanie danymi badawczymi



Praca w grupach: plan RDM

- Proszę wybrać przykładowy projekt badawczy
- Jakie dane zostaną wytworzone?
- Które dane zachować i udostępnić?

Gdzie szukać otwartych danych?

- Repozytoria
- Biblioteki cyfrowe (pytanie o API) i Europeana (agregator)
- Czasopisma o danych

Czasopisma publikujące dane (*data journals*)

SCIENTIFIC DATA

npg nature publishing group

ELSEVIER Data in Brief

- Artykuły opisujące dane (*data descriptors*)
- Dane są deponowane w repozytoriach
- Niektóre czasopisma dopuszczają też możliwość dołączania danych w postaci Supplementary Material

→ Uzupełnienie systemu repozytoryjnego, nie alternatywa

Journal of Open
Humanities Data

ubiquity press open access

About this Journal

The *Journal of Open Humanities Data* (JOHD) features peer reviewed publications describing humanities data or techniques with high potential for reuse. Humanities subjects of interest to JOHD include, but are not limited to Art History, History, Linguistics, Literature, Music, Philosophy, Religious Studies, etc. Data that crosses one or more of these traditional disciplines are highly encouraged.

LATEST ARTICLES

Article Processing Charges Paid by 25 UK Universities in 2014

Lawson — 29 Sep 2015

Share:

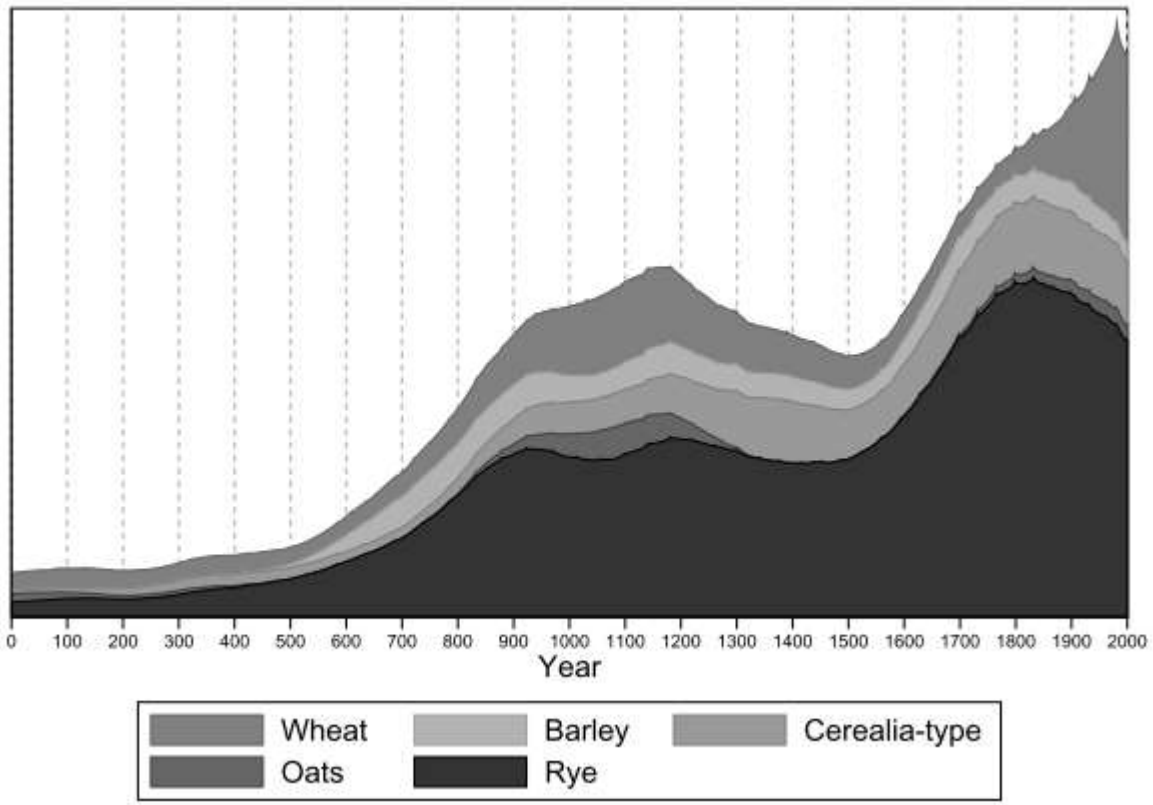
Vagrant Lives: 14,789 Vagrants Processed by the County of Middlesex, 1777–1786

Crymble et al. — 29 Sep 2015

Share:

A gdyby tak poszukać danych poza humanistyką?

Figure 2: Agricultural Output in Greater Poland, 0–2000 AD



On the Use of Palynological Data in Economic History: New Methods and an Application to Agricultural Output in Central Europe, 0–2000 AD*

A. Izdebski† G. Koloch‡ T. Słoczyński§

M. Tycner-Wolicka¶

Abstract: In this paper we introduce a new source of data to economic history: palynological data, i.e. information about pollen grains which are preserved in bottom sediments of various water basins. We discuss how this data is collected and how it should be interpreted; develop new methods for aggregating this information into regional trends in agricultural output; construct an extensive data set with a large number of pollen sites from Central Europe; and use our methods to study the economic history of Greater Poland, Lesser Poland, Bohemia, Brandenburg, and Lower Saxony since the first century AD.

Open Data Button <beta>

Take action Requests About Blog Login Download

Open Data Button beta

ndatabutton.org/action

Need access to research data?

We can help.

Push Button. Request Data. Make Progress.

Hidden data is hindering research. The Open Data Button allows you to request research data at the click of a button. When a researcher releases the requested data it can be

HOW THE BUTTON WORKS

Push Button.

Download the Open Data Button for your browser. Next time you're reading a research paper and you want to investiage the data behind it, push the Open Data Button.

Request Data.

The Open Data Button will try to find you the data you need. If that doesn't work, it'll start a request to the author asking them to share their data. Authors will be able to share their data with you and the world.

Make Progress.

When you get the data you need, you can make progress with your work. Your story will be used to encourage researchers to make their data available, making progress to more transparent and reproducible research.

Podsumowanie

- Udostępnienie danych badawczych => korzyści dla nauki i naukowca
- Warto planować RDM na początku projektu
- Plan zarządzania danymi badawczymi

Przydatne linki:

- Pon.edu.pl
- Otwartanauka.pl http://otwartanauka.pl
- https://repod.pon.edu.pl/pl/

- Digital Curation Centre http://www.dcc.ac.uk/
- CODATA http://www.codata.org/
- OpenAIRE https://www.openaire.eu/

Dziękujemy za uwagę

Kontakt:

msommer@icm.edu.pl

m.starczewski@icm.edu.pl

UNIWERSYTET WARSZAWSKI
Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego
www.icm.edu.pl

PLATFORMA OTWARTEJ NAUKI