Zarządzanie danymi badawczymi

Marta Hoffman-Sommer

Platforma Otwartej Nauki, ICM, Uniwersytet Warszawski

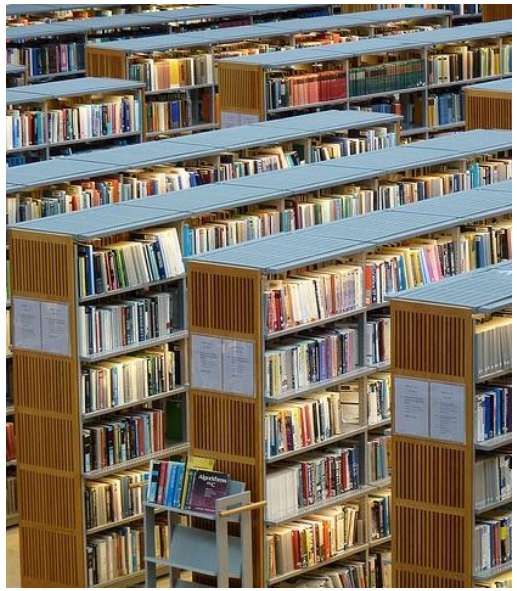
Warszawa, 22.04.2015

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

PLATFORMA OTWARTEJ NAUKI

Co to są dane badawcze?

Rezultaty badań naukowych



KTH Biblioteket, CC-BY-SA
https://www.flickr.com/photos/kthbiblioteket/44726400423/

Publikacje naukowe
(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

PLATFORMA OTWARTEJ NAUKI

Różne definicje danych badawczych

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

„Dane badawcze to dane zebrane, zaobserwowane lub wytworzone jako materiał do analizy, w celu uzyskania oryginalnych wyników naukowych.”

„…’dane badawcze’ definiujemy jako zapisy faktów (wartości liczbowe, zapisy tekstowe, obrazy i dźwięki), które służą jako źródła pierwotne w badaniach naukowych, i które są powszechnie uznawane przez społeczność naukową za niezbędne do oceny wyników naukowych. Zbiór danych badawczych stanowi usystematyzowaną, częściową reprezentację badanego zjawiska.”

Co zaliczamy do danych badawczych?

Dokumenty tekstowe, notatki

Dane liczbowe

Kwestionariusze, ankiety, wyniki badań ankietowych

Nagrania audio i video, zdjęcia

Próbki, artefakty, obiekty

Zawartość baz danych (video, audio, teksty, obrazy)

Modele matematyczne, algorytmy

Oprogramowanie (skrypty, pliki wejściowe…)

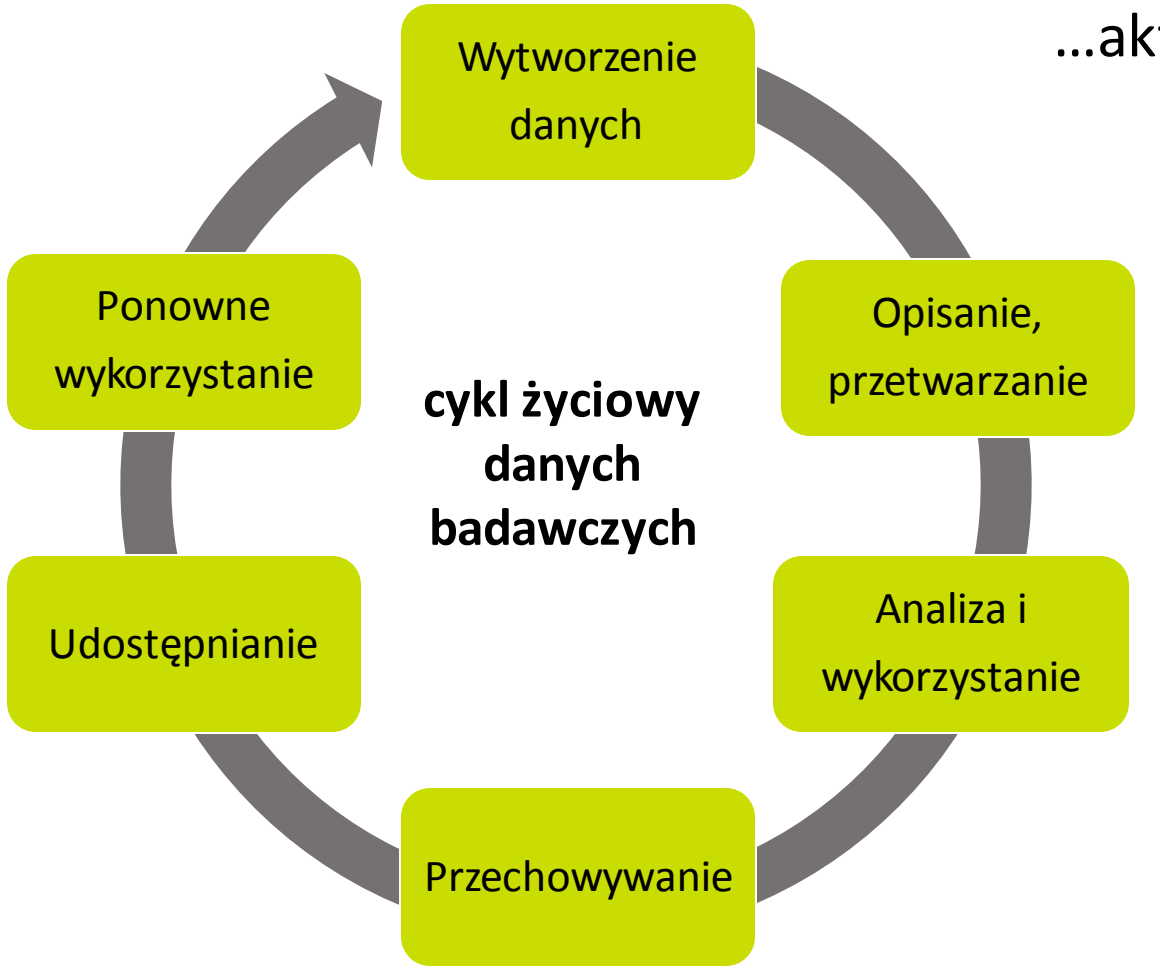
Wyniki symulacji komputerowych

Protokoły laboratoryjne, opisy metodologiczne

Co to jest zarządzanie danymi badawczymi?

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

PLATFORMA OTWARTEJ NAUKI



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

Jakie aspekty należy uwzględnić w ZDB?

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

Jakie wymagania dotyczące zarządzania danymi mogą nas dotyczyć?

- Wymagania nakładane przez **wydawców naukowych**: konieczność udostępniania danych.

  Nature, PLoS, Amer. Economic Review…

- Wymagania nakładane **w umowach grantowych**: konieczność tworzenia Planów Zarządzania Danymi i/lub udostępniania danych.

  Komisja Europejska, brytyjskie RCUK, amerykańska NSF…

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**.”

icm

PLATFORMA OTWARTEJ NAUKI

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.”

„Na ile to możliwe, projekty są zobowiązane do podjęcia działań umożliwiających osobom trzecim dostęp do danych badawczych, ich analizę maszynową, ponowne wykorzystanie, kopiowanie i rozpowszechnianie (bez opłat ze strony użytkowników).

Prostą i skuteczną metodą osiągnięcia powyższego celu jest dołączenie do deponowanych danych licencji Creative Commons (CC-BY lub oświadczenia CC0).”

- Pilotaż obejmuje projekty z 7 wybranych obszarów tematycznych.
- Można się włączyć do pilotażu (*opt-in*), można się też wyłączyć (*opt-out)*.

Kiedy można się wyłączyć z pilotażu?

- Gdy planowane jest **komercyjne lub przemysłowe** wykorzystanie danych
- Gdy uczestnictwo stoi w sprzeczności z wymogami poufności, związanymi z **bezpieczeństwem**
- Gdy stoi w sprzeczności z obowiązującymi zasadami **ochrony danych osobowych**
- Gdyby udział w pilotażu uniemożliwiał **osiągnięcie głównego celu** naszych działań
- Jeżeli w ramach projektu **nie zostaną wytworzone ani zebrane** żadne dane naukowe
- Gdy występują inne **uzasadnione przyczyny** by nie uczestniczyć w pilotażu

Można się wyłączyć zarówno na etapie wniosku grantowego, jak i w trakcie trwania projektu. Powody wyłączenia należy wyjaśnić w Planie Zarządzania Danymi.

Zarządzanie danymi badawczymi

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

PLATFORMA OTWARTEJ NAUKI

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 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)?

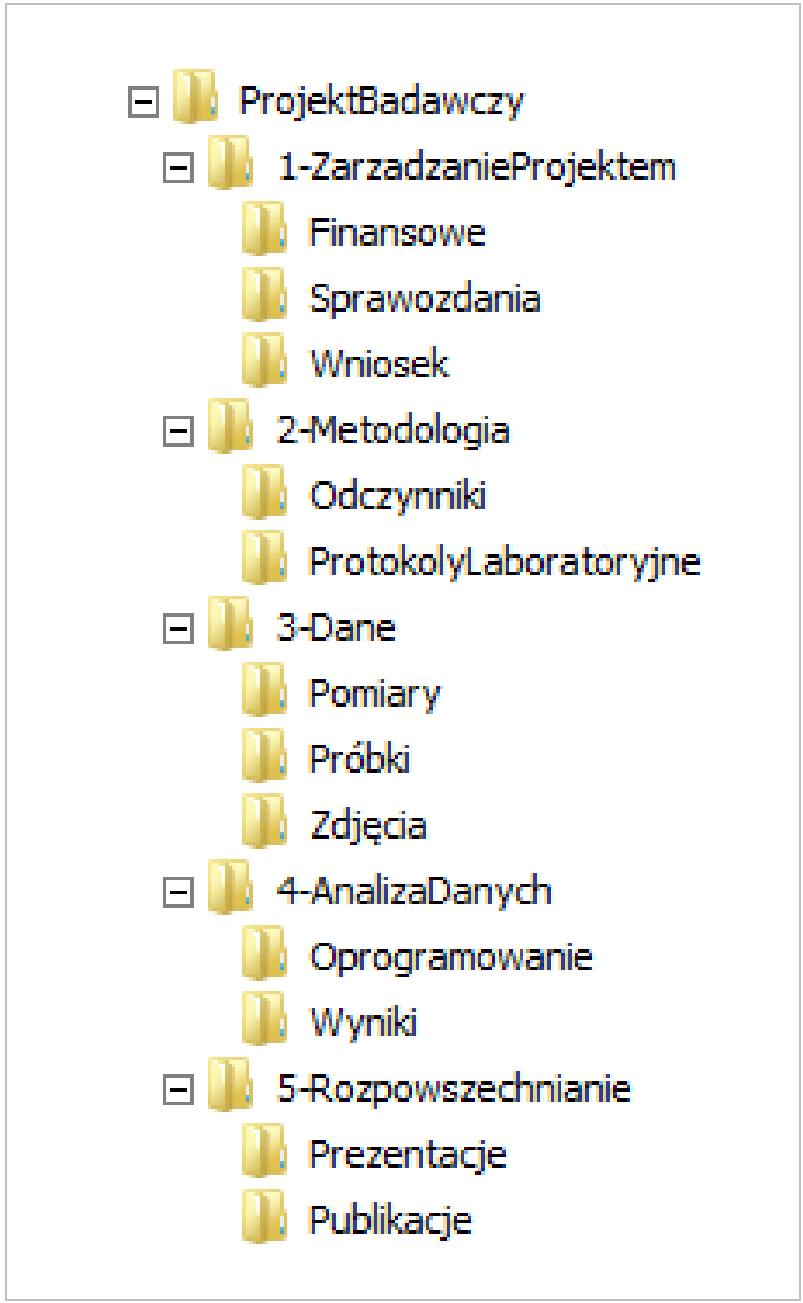
Na podstawie: Workbook for Writing a Data Management Plan, http://www.dcc.ac.uk/training/digital-curation-101/dmp-workshop-uct

Struktura i nazwy folderów

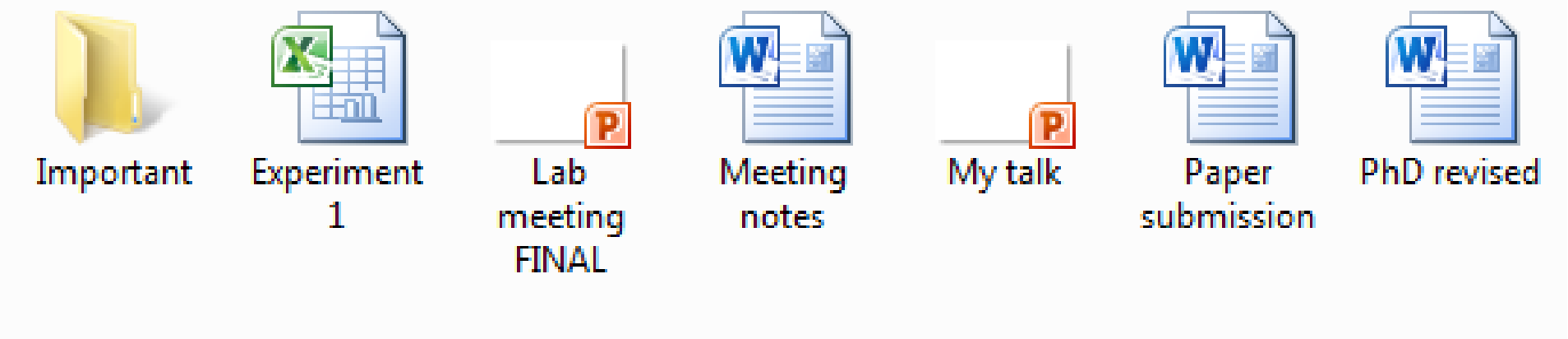
- Hierarchiczna struktura – nie za głęboka, nie za szeroka
- Możliwe też tagowanie plików

Różne przykłady:
http://www.vukovicnikola.info/folder-structure-for-research/
http://pgbovine.net/research-directory-structure.htm
https://www.ukdataservice.ac.uk/manage-data/format/organising

Przykład



Nazewnictwo plików – po co się tym zajmować?



Czy za 3 lata będziemy wiedzieli, co jest w tych plikach?

Nazewnictwo plików

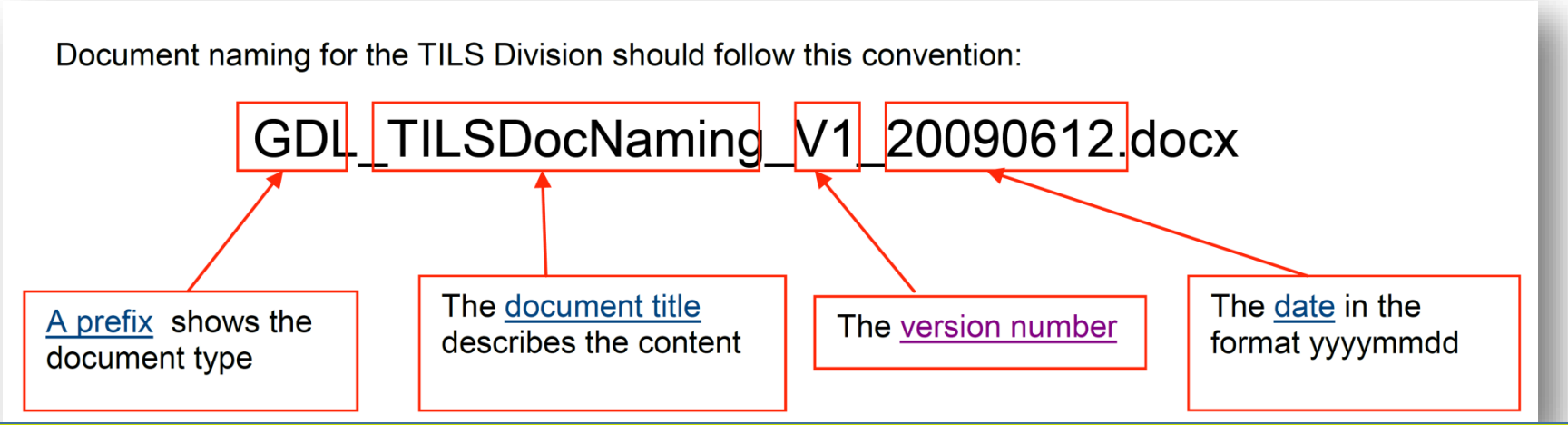
- Jakie informacje potrzebujemy zawrzeć w nazwie?

  (rodzaj danych, inicjały badacza, numer próbki, data, numer wersji, etc.)

- Czy nasze nazwy są unikalne?
- Jak będziemy chcieli sortować pliki?

http://www.data.cam.ac.uk/files/gdl_tilsdocnaming_v1_20090612.pdf

Przykład:



Informujemy gdzieś w dokumentacji co oznaczają nasze nazwy.

Co najmniej 2 backupy, w tym jeden *off-site*:

każdy poniedziałek rano

codziennie rano
- automatycznie

- Regularność
- Automatyzacja

Na podstawie slajdu:
Y. Creba, V. Philips, C. Sewell, M. Teperek,
University of Cambridge, CC-BY

Darmowe oprogramowanie do zarządzania backupami (przykładowe):
http://www.2brightsparks.com/download-syncbackfree.html

Więcej na temat bieżącego zarządzania:

http://libraries.mit.edu/data-management/services/workshops/

a w szczególności o organizowaniu danych:

http://libraries.mit.edu/data-management/files/2014/05/FileOrg_20160121.pdf

albo tu:

https://www.ukdataservice.ac.uk/manage-data/format/organising

http://datalib.edina.ac.uk/mantra/organisingdata/

https://library.stanford.edu/research/data-management-services/data-best-practices/best-practices-file-naming

Narzędzia ułatwiające bieżące zarządzanie danymi

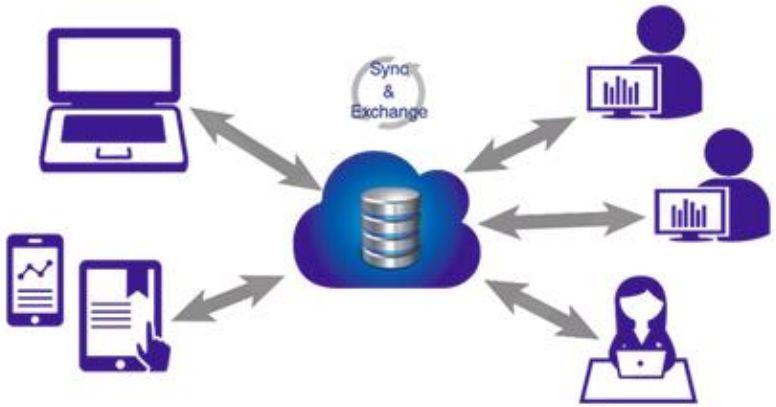
1. **EUDAT** – projekt europejski tworzący otwartą dla wszystkich e-infrastrukturę, m.in. do przechowywania danych badawczych i zarządzania nimi: **B2DROP**.
2. **OSF.io** – serwis stworzony przez amerykańską organizację *non-profit* Center for Open Science – służy do bieżącego zarządzania prowadzonymi badaniami. Pozwala na integrację z innymi serwisami, np. Dropbox, Google Drive, GitHub, FigShare, Mendeley.
3. **GitHub** – przechowywanie i wersjonowanie oprogramowania; wykorzystywane również dla innych rodzajów plików.

B2DROP

B2DROP is a secure and trusted data exchange service for researchers and scientists to keep their research data synchronized and up-to-date and to exchange with other researchers.

An ideal solution to:

- store and exchange data with colleagues and team members,
- synchronise multiple versions of data,
- ensure automatic desktop synchronisation of large files.



Users can

- define with whom to exchange data, for how long and how
- are offered up to 20GB of storage space for research data
- access and manage permissions to files from any device and any location.

Use B2DROP | Contact

Services

- B2ACCESS
- B2DROP
- B2SHARE
- B2STAGE
- B2SAFE
- B2FIND

https://www.eudat.eu/services/b2drop

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

PLATFORMA OTWARTEJ NAUKI

Dane, których nie zamierzamy przechowywać

Dokumentować:

Co, dlaczego i kiedy zostało wyrzucone

Jakość danych

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

Ćwiczenie: Identyfikacja i selekcja danych

Wybieramy w grupie jeden projekt naukowy i dla tego projektu:

1. Określamy, jakie dane naukowe zostaną wytworzone lub zebrane (wszystkie!)
2. Zastanawiamy się, które spośród tych danych warto archiwizować (które chcemy przechowywać po zakończeniu projektu)

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

Archiwizacja: przechowywanie długoterminowe

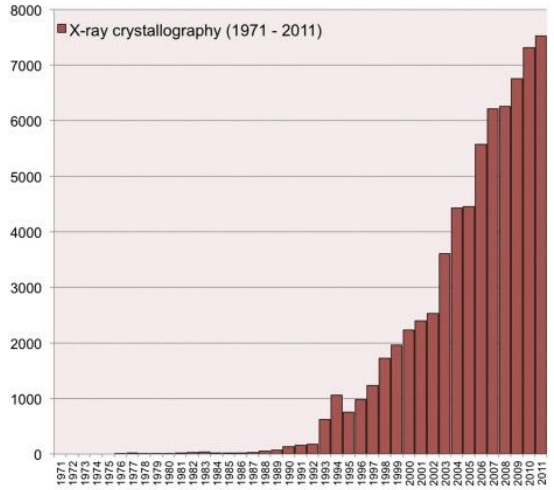
1. Bezpieczeństwo danych – repozytorium/archiwum godne zaufania
2. Widoczność – znane wśród badaczy, dobrze widoczne w wyszukiwarkach
3. Trwała lokalizacja – stały identyfikator cyfrowy (np. DOI – *digital object identifier*)

Cyfrowe repozytoria danych badawczych

- specjalistyczne (wąskie dziedzinowo)
- instytucjonalne
- szeroko zakrojone tematycznie
- ogólne

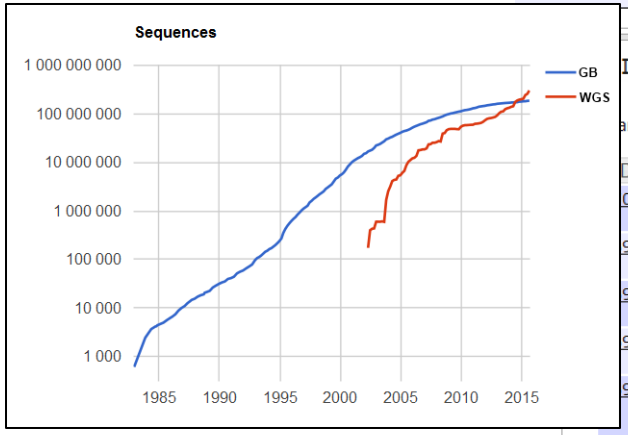
Repozytoria specjalistyczne

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

Repozytorium danych społecznych prowadzone przez ISS UW i IFIS PAN

Repozytoria instytucjonalne

Purdue University Research Repository

Repozytorium uczelniane

BODC British Oceanographic Data Centre NATURAL ENVIRONMENT RESEARCH COUNCIL

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

icm

PLATFORMA OTWARTEJ NAUKI

Repozytoria ogólne

Krajowe repozytorium danych: Holandia

Repozytorium ogólnodostępne (publikacje + dane)

Krajowe repozytorium danych: Polska

Repozytorium ogólnodostępne (publikacje + dane)

Czasopisma publikujące dane (*data journals*)

SCIENTIFIC DATA

npg nature publishing group

ELSEVIER Data in Brief

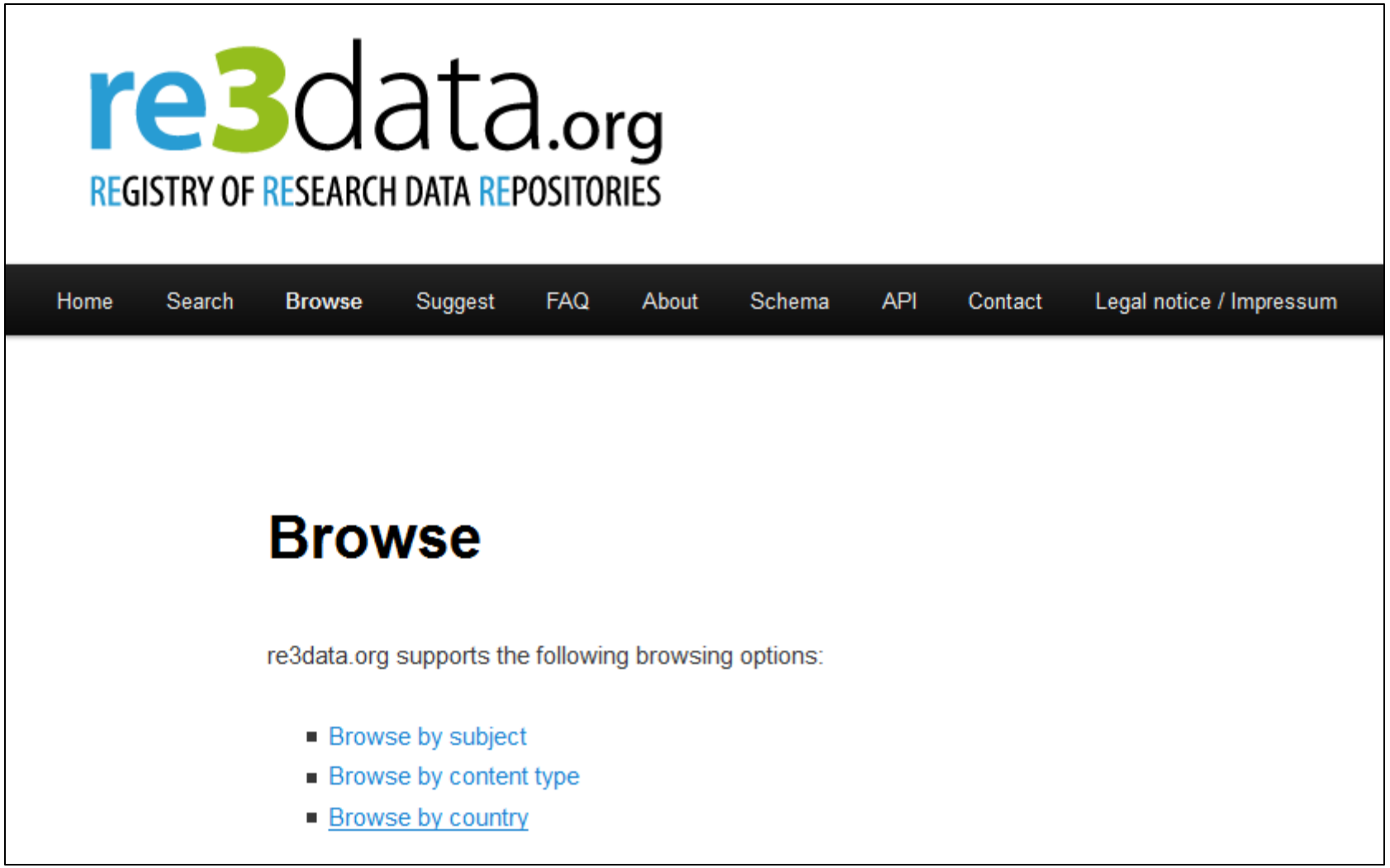
- Artykuły opisujące dane (*data descriptors*)
- Dane są deponowane w repozytoriach
- Niektóre czasopisma dopuszczają też możliwość dołączania danych w postaci Supplementary Material

→ Uzupełnienie systemu repozytoryjnego, nie alternatywa

icm

PLATFORMA OTWARTEJ NAUKI

re3data.org – wyszukiwarka repozytoriów



4. Przygotowanie danych

Przygotowanie plików (ew. anonimizacja danych)

Metadane

Dokumentacja

Dobór formatów plików do archiwizacji

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

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

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

Metadata standards: na stronach

Digital Curation Centre

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

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

Dokumentacja: na poziomie projektu

- opis celu i kontekstu badań
- opisy metodologiczne sposobów pozyskania danych
- struktura plików z danymi, relacje między plikami
- linki do publikacji powiązanych z danymi
- formularze zgody na udział w badaniu
- zastosowane metody kontroli lub walidacji danych
- informacje o różnych wersjach zbiorów danych
- informacje o zbiorach danych, których z przyczyn prawnych nie można udostępnić

Na podstawie: http://www.data-archive.ac.uk/create-manage/document/data-level

Dokumentacja: na poziomie zbioru danych

- nazwy i opisy zastosowanych zmiennych, rekordów, wartości; jednostki pomiarowe
- opisy zastosowanych schematów klasyfikacyjnych
- informacje o urządzeniach pomiarowych (linki), informacje o zastosowanych ustawieniach, metodach kalibracji
- oznaczenia brakujących wartości oraz wyjaśnienie, dlaczego ich brakuje
- dane przetworzone, wraz z oprogramowaniem lub algorytmami zastosowanymi do ich uzyskania
- lista wszystkich opisanych przypadków lub obiektów (dla badań jakościowych)
- jeżeli sprawdzamy w jakikolwiek sposób jakość naszych danych: wyniki tych testów

Na podstawie: http://www.data-archive.ac.uk/create-manage/document/data-level

Gdzie umieszczamy dokumentację?

→ Wpisana w same pliki z danymi

np. informacja o jednostkach pomiarowych w nagłówku kolumny w tabeli

→ Dołączona do zbioru danych w osobnym pliku / plikach

często w postaci pliku ReadMe.txt

(jeden dla całego zbioru danych albo po jednym dla każdego pliku)

Guidelines to writing

„readme" style metadata

Autorka: Wendy Kozlowski,
Research Data Management Service Group,
Cornell University Libraries

http://data.research.cornell.edu/sites/default/files/SciMD_ReadMe_Guidelines_v4_1_0.pdf

1. *Introductory information*
    a. **For each filename, a short description of what data it contains**
    b. Format of the file if not obvious from the file name
    c. If the data set includes multiple files that relate to each other, the relationship between the files or a description of the file structure that holds them (possible terminology might include "dataset" or "study" or "data package")
    d. **Name/institution/address/email information for**
        i. **Principle investigator (or person responsible for collecting the data)**
        ii. Associate or co-investigators
        iii. Contact person for questions
    e. **Date of data collection (can be a single date, or a range)**
    f. **Information about geographic location of data collection**
    g. **Date that the file was created**
    h. Date(s) that the file(s) was updated and the nature of the update(s), if applicable
    i. Keywords used to describe the data topic
    j. Language information

2. *Methodological information*
    a. **Method description, links or references to publications or other documentation containing experimental design or protocols used in data collection**
    b. Any instrument-specific information needed to understand or interpret the data
    c. Standards and calibration information, if appropriate
    d. Describe any quality-assurance procedures performed on the data
    e. Definitions of codes or symbols used to note or characterize low quality/questionable/outliers that people should be aware of
    f. People involved with sample collection, processing, analysis and/or submission

3. *Data-specific information*
    a. **Full names and definitions (spell out abbreviated words) of column headings for tabular data**
    b. **Units of measurement**
    c. **Definitions for codes or symbols used to record missing data**
    d. **Specialized formats or abbreviations used**

4. *Sharing/Access information*
    a. Licenses or restrictions placed on the data[2]
    b. Links to publications that cite or use the data
    c. Links to other publicly accessible locations of the data
    d. **Recommended citation for the data**
    e. Information about funding sources that supported the collection of the data

Dane osobowe

- **Anonimizacja danych** – żeby niemożliwe było zidentyfikowanie uczestników badania
- lub: **Zgoda osób badanych** na udostępnienie / nadanie licencji

Darmowe narzędzie do anonimizacji danych badawczych (niegotowe): https://www.openaire.eu/anonymizing-your-data

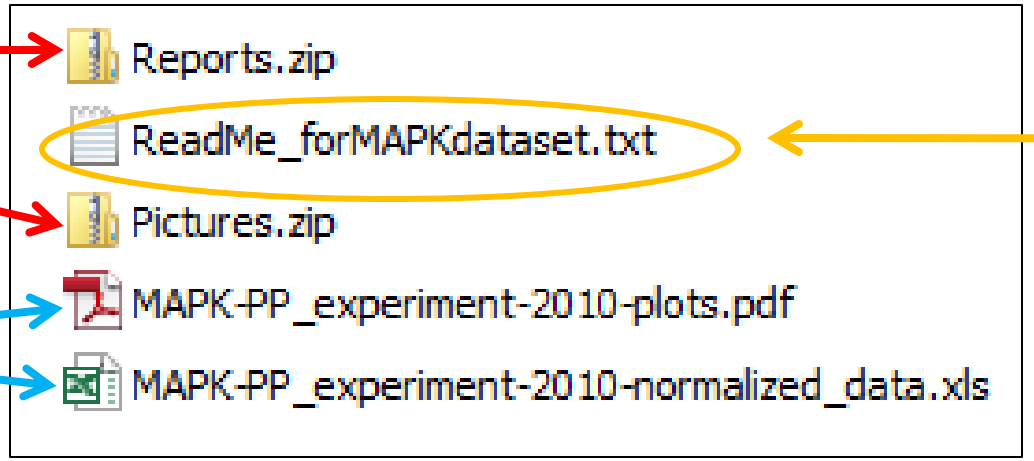
- Być może musimy wprowadzić ograniczenia dostępu – trzeba sprawdzić, czy wybrane przez nas repozytorium jest w stanie je zaimplementować

5. Deponowanie danych

Surowe dane: .txt

Dane przetworzone: .jpeg

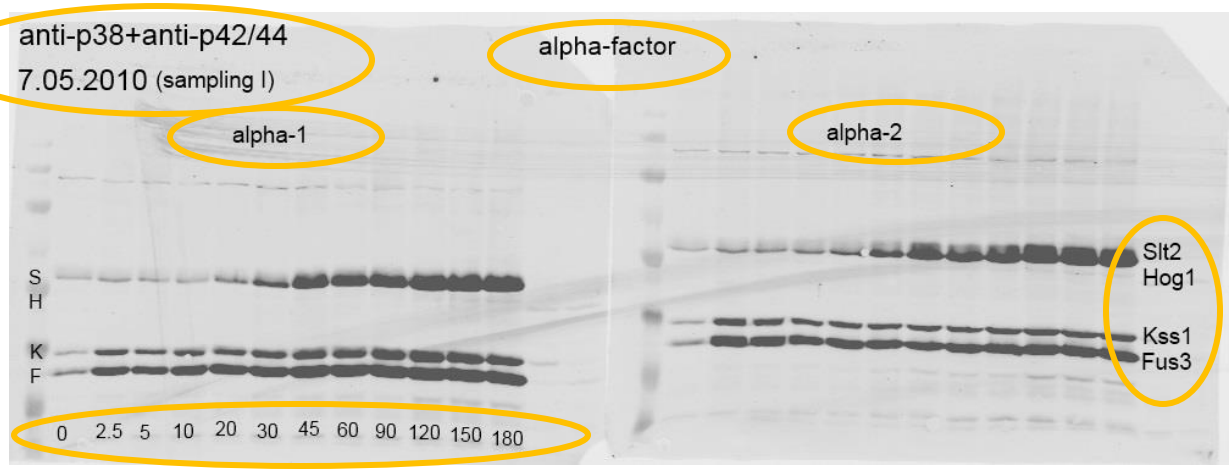
Analizy danych:
.xls, .pdf



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

icm

PLATFORMA OTWARTEJ NAUKI

RepOD - serwis dla polskiej społeczności akademickiej

repod.pon.edu.pl

Dane:

(1) badawcze

(2) otwarte

→ ze wszystkich dziedzin nauki

→ wszystkie formaty plików

Zbiór danych

(1) metadane

Time-course data of the phosphorylation of MAP kinases in yeast cells treated with alpha-factor and/or salt

This dataset contains the results of an experiment meant to investigate crosstalk between Ste11-dependent mitogen-activated protein kinase (MAPK) pathways in cells of the yeast Saccharomyces cerevisiae. Yeast cells were stimulated with the hormone alpha-factor and with salt (NaCl) in three different scenarios: (1) only alpha-factor, (2) alpha-factor and salt simultaneously, (3) only salt. Samples were drawn from the cultures at defined time points after stimulation, total proteins were isolated and analyzed by Western blotting with antibodies against the phosphorylated forms of three yeast MAPKs: Fus3, Kss1 and Hog1. The Western blots were scanned and quantified using an Odyssey infrared scanner. This dataset contains raw quantitative phosphorylation data, processed data (including graphs), and scans of the blots.

Publisher: RepOD

Publication year: 2015

Type of resource: Collection

Area of study: Physical and mathematical sciences

Funder: European Commission

Funding program: EC-FP6

Grant number: 043310

License for files: CC0-1.0

Keywords

HOG pathway | MAP kinase | S. cerevisiae | cell signalling | crosstalk | pheromone pathway | quantitative Wester...

Authors

Author	Affiliation
Hoffman-Sommer, Marta	Theoretical Biophysics, Institute of Biology, Humboldt University, Germany, and Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Poland
Klaus, Christine	Theoretical Biophysics, Institute of Biology, Humboldt University, Germany
Klipp, Edda	Theoretical Biophysics, Institute of Biology, Humboldt University, Germany

(2) pliki

Pliki w tym zbiorze

ReadMe_forMAPKdataset
TXT
This is a text file containing all information necessary to understand and...
6 wejść | 1 pobranie
Eksploruj

Reports
ZIP
This archive contains text files (tab-separated values) containing the raw...
1 wejście | 2 pobrania
Eksploruj

Pictures
ZIP
This archive contains jpeg files presenting scans of all Western blots...
0 wejść | 3 pobrania
Eksploruj

MAPK-PP-normalized_data
XLS
This is an xls spreadsheet containing processed quantitative data. Data...
4 wejścia | 1 pobranie
Eksploruj

MAPK-PP-plots
PDF
This file contains images of graphs obtained by plotting the processed data...
6 wejść | 9 pobrań
Eksploruj

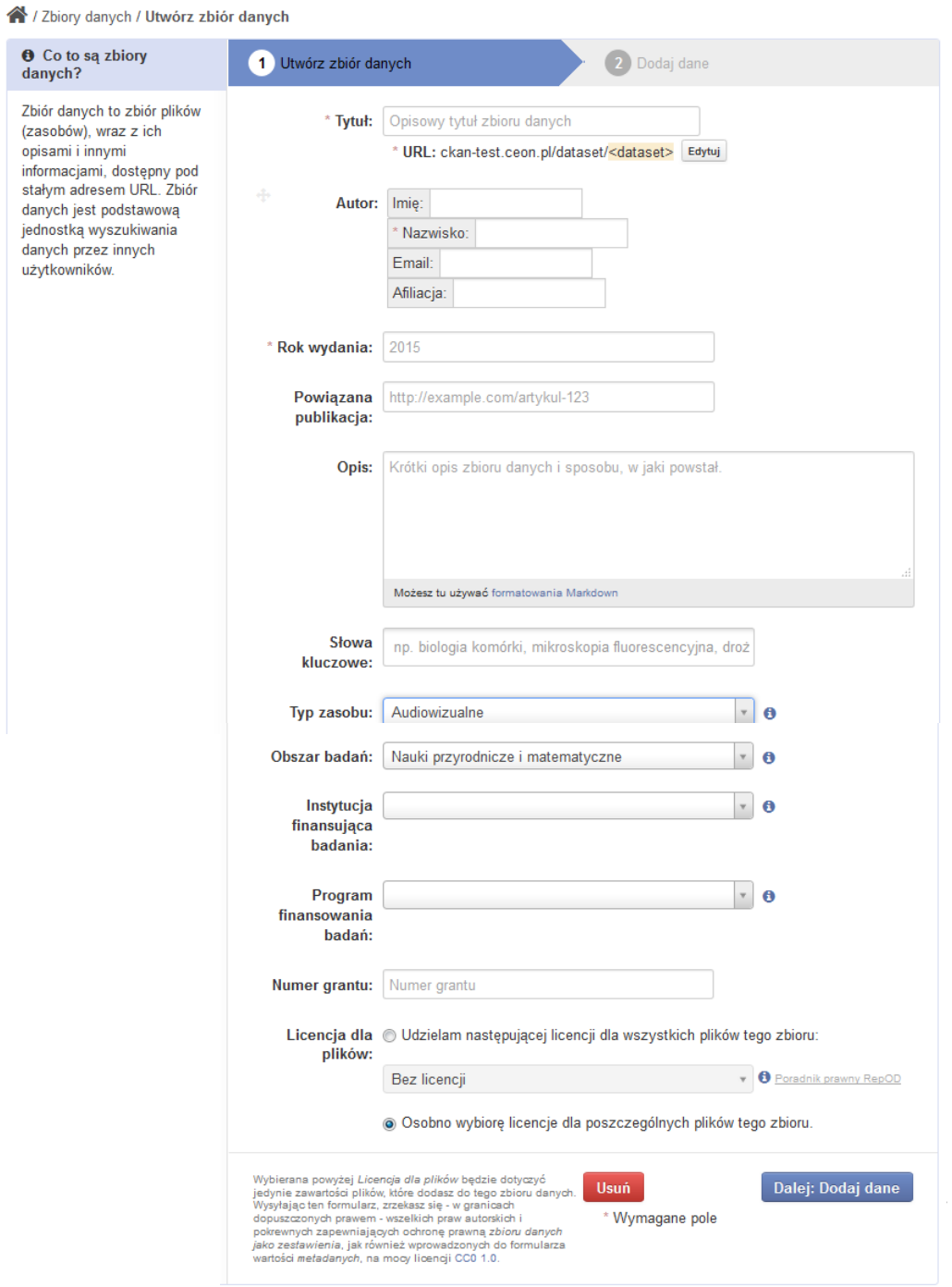
Metadane

- Tytuł, autor(autorzy), rok wydania, krótki opis…
- Link do powiązanej publikacji naukowej
- Informacje o finansowaniu badań
- Licencja prawna

Repozytorium przydziela numer

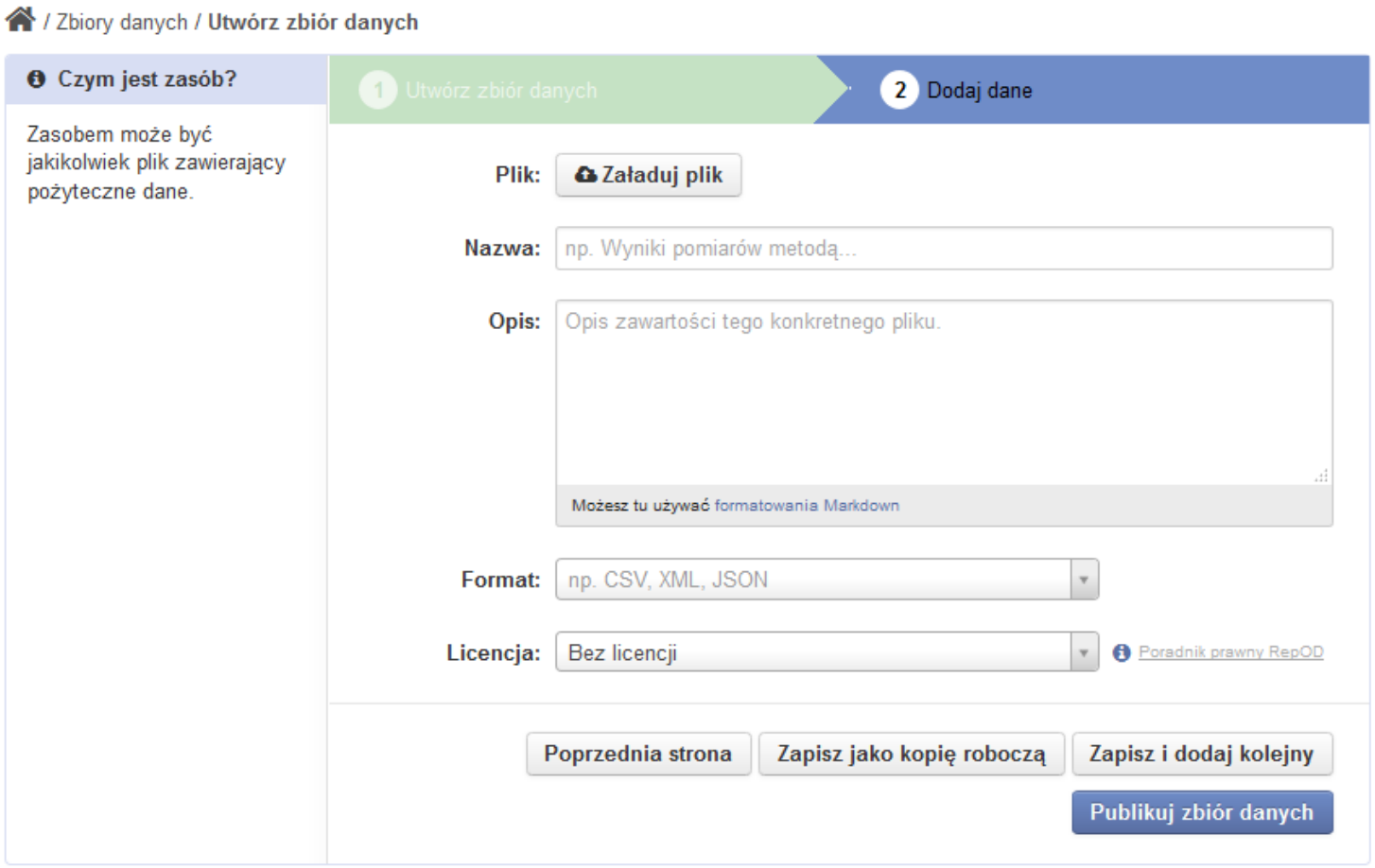
DOI – *digital object identifier*

→ cytowanie danych



Pliki

- Dowolnie wiele plików w zbiorze
- Różnorodność formatów



PLATFORMA OTWARTEJ NAUKI

Kwestie prawne

- Kto ma prawo decydować o udostępnieniu danych badawczych?

  → Komu przysługują prawa do zbioru danych?

- Jakie inne ograniczenia prawne nas obowiązują przy udostępnianiu danych?

- Co wolno użytkownikowi danych? Kto i w jaki sposób o tym decyduje?

  → Jakie prawa przysługują użytkownikowi na podstawie zasad dozwolonego użytku i jak można je rozszerzyć za pomocą licencji?

Prawa związane ze zbiorem danych badawczych

- **Prawa autorskie** – związane z **kreatywnością** wykonanej pracy (fakty nie podlegają ochronie prawnoautorskiej)
- **Prawa do baz danych** – związane z **inwestycją** włożoną w utworzenie danej bazy (w krajach UE)
- **Prawa osób trzecich** – związane z **innymi osobami** zaangażowanymi w powstanie danych (twórcy zawartości, uczestnicy badania, …)

- Różne systemy prawne w różnych krajach
- Może się zdarzyć tak, że pewne prawa do danych naukowych przysługują pracodawcy („Regulamin nabywania, korzystania i ochrony własności intelektualnej” uczelni)

Co trzeba ustalić, by udostępnić dane?

Musimy ustalić, kto dysponuje prawami do zbioru danych:

1. Kontaktujemy się ze (współ)autorami.
2. Sprawdzamy regulacje między autorami a pracodawcą.
3. Rozważamy, czy występują prawa osób trzecich.

Uwzględniamy inne ograniczenia prawne:

1. Czy występują dane osobowe? anonimizacja
2. Na co wyrazili zgodę uczestnicy badania? formularz zgody
3. Inne sytuacje, gdy nie wolno nam publikować danych:

bezpieczeństwo kraju, gatunki chronione, tajemnica przedsiębiorstwa, etc.

Prawny status udostępnionych danych

Dane w repozytorium możemy udostępnić:

- bez licencji: na zasadach dozwolonego użytku
- z licencją (np. na otwartej licencji Creative Commons)
- z oświadczeniem o zrzeczeniu praw (np. Creative Commons Zero)

Dozwolony użytek

…obowiązuje automatycznie – na mocy ustawy – gdy nie dołączymy żadnych dodatkowych oświadczeń prawnych.

Licencjonowanie

- Zalecane jest korzystanie ze **standardowych licencji**
- **Międzynarodowe licencje** – tak konstruowane, by (na ile się da) prowadziły do tych samych skutków w różnych systemach prawnych
- Od restrykcyjnych po bardzo liberalne (otwarte licencje)

Pilotaż Otwartych Danych w H2020

„Na ile to możliwe, projekty są zobowiązane do podjęcia działań umożliwiających osobom trzecim dostęp do danych badawczych, ich analizę maszynową, ponowne wykorzystanie, kopiowanie i rozpowszechnianie (bez opłat ze strony użytkowników).

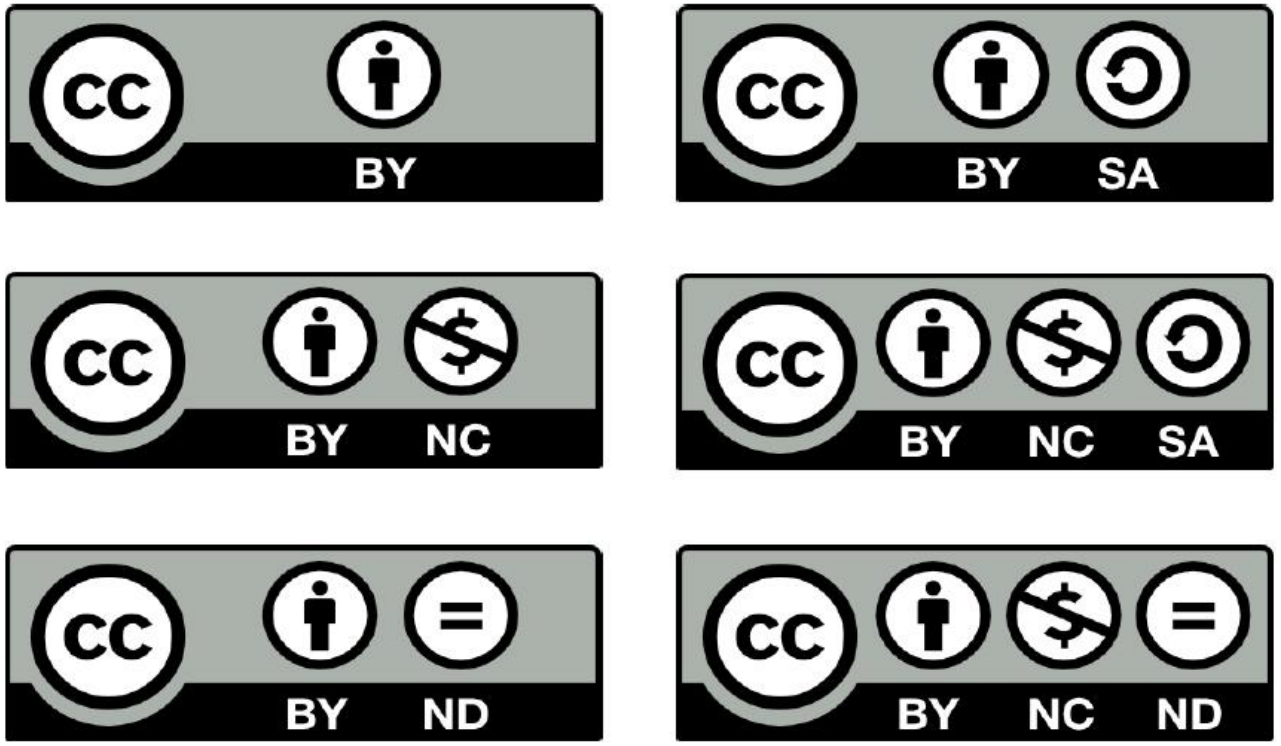
Prostą i skuteczną metodą osiągnięcia powyższego celu jest dołączenie do deponowanych danych licencji Creative Commons (CC-BY lub CC0).”

?

icm

PLATFORMA OTWARTEJ NAUKI

Co to są licencje Creative Commons (CC)?



Co oznaczają ikony licencji CC?

BY – Attribution

/Uznanie autorstwa

SA – Share Alike

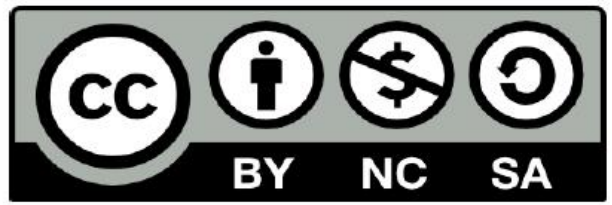
/Na tych samych warunkach

NC – Non-commercial

/Użycie niekomercyjne

ND – No derivatives

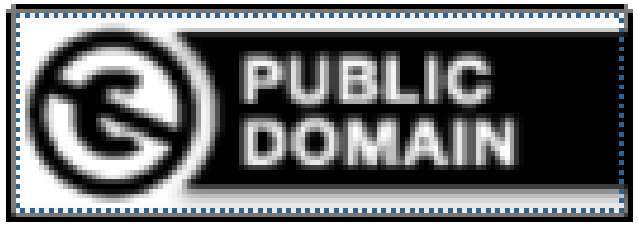
/Bez utworów zależnych



icm

PLATFORMA OTWARTEJ NAUKI

Domena publiczna – dzieła nieobjęte ochroną związaną z autorskimi prawami majątkowymi



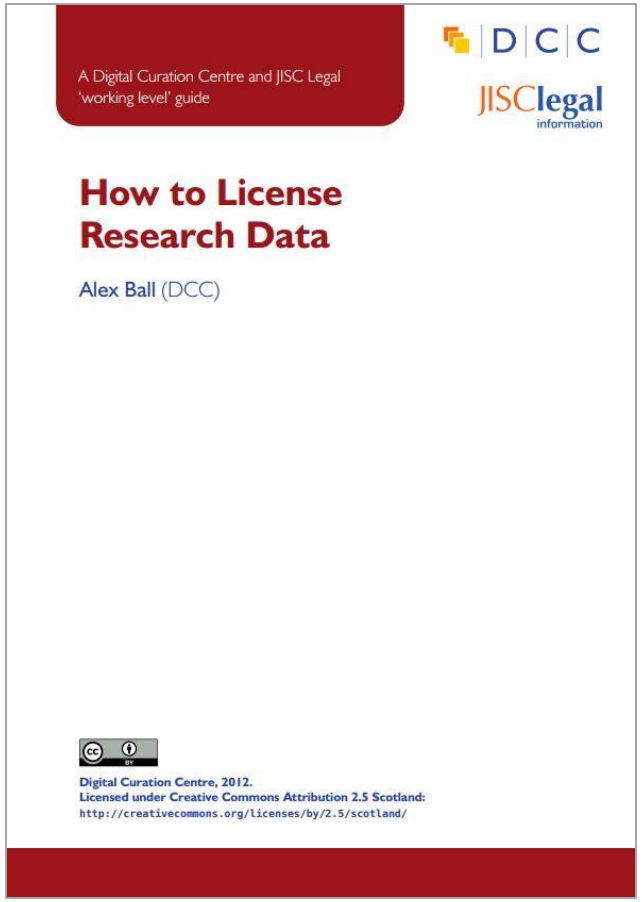
Znak domeny publicznej

gdy wiemy, że dzieło nie jest chronione

Przekazanie do domeny publicznej

gdy chcemy, by nasze dzieło nie było chronione

Jak licencjonować dane – poradnik



www.dcc.ac.uk/resources/how-guides/license-research-data

PLATFORMA OTWARTEJ NAUKI

Dlaczego CC0 jest zalecane dla danych?

BY: Zbiory danych są szczególnie podatne na tzw. *nawarstwianie atrybucji* (*attribution stacking*) – gdy w zbiorze pochodnym musimy zaznaczyć autorstwo wszystkich danych, z których skorzystaliśmy, nawet bardzo już odległych.

SA: Licencje „Na tych samych warunkach” (tzw. *copyleft*) utrudniają łączenie zbiorów danych z innymi zbiorami, które mogą mieć inne licencje typu SA.

NC: Niejasne: według niektórych interpretacji prawnych licencje niekomercyjne nie pozwalają nawet na wykorzystanie do wytworzenia artykułu naukowego, który jest potem sprzedawany (jako część czasopisma naukowego).

Na podstawie:
Ball, A. (2014). ‘How to License Research Data’.
DCC How-to Guides. Edinburgh: Digital Curation Centre. Dostępne online:
http://www.dcc.ac.uk/resources/how-guides/license-research-data#x1-4000

icm

PLATFORMA OTWARTEJ NAUKI

CC0 jest łatwe w użyciu

- Nie musimy ustalać, jakie prawa i do jakich elementów zbioru nam przysługują (autorskie, do baz danych).
- Bierzemy pod uwagę tylko prawa osób trzecich.

Czy wszystkie dane powinny być otwarte? Nie.

Dane osobowe

Bezpieczeństwo narodowe

Komercjalizacja wyników badań

Ochrona gatunków zagrożonych, itp.

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

Licencjonowanie

- Wybieramy właściwą licencję
- Upewniamy się, że mamy prawo ją zastosować
- Upewniamy się, że repozytorium, w którym zamierzamy zdeponować nasze dane, przyjmie tę licencję
- Umieszczamy informację o licencji w widocznym miejscu w dokumentacji oraz informujemy o niej podczas deponowania

Po co udostępniać dane?

…aby umożliwić weryfikację wyników naukowych.

Ćwiczenie: Udostępnianie danych

Zakreślamy 3 największe przeszkody powstrzymujące nas przed udostępnianiem danych. Dlaczego są tak istotne?

A w naszym projekcie, co udostępnimy i na jakich zasadach?

Moje dane zawierają dane osobowe lub wrażliwe	Moje dane są zbyt skomplikowane	Inni użytkownicy mogą błędnie zinterpretować moje dane	Moje dane nie są ciekawe
Moje badania są finansowane przez firmę komercyjną, która nie zgadza się na ich udostępnienie	Może jeszcze będziemy chcieli skorzystać z tych danych w innym artykule	Ludzie będą się ze mną kontaktować i zadawać mi różne pytania	Dane są poufne/ Wpływają na bezpieczeństwo kraju
Moje dane są za duże	Inni zobaczą, że moje dane są kiepskiej jakości	Chcę opatentować mój wynalazek	Jestem bardzo zajęta/y i to nie jest mój priorytet
Nie wiem jak to zrobić	Nie jestem pewna/y, kto jest właścicielem moich danych	Ktoś mógłby ukraść lub splagiatować moje dane	Instytucja finansująca moje badania tego nie wymaga

PLATFORMA OTWARTEJ NAUKI

Plan Zarządzania Danymi

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

PLATFORMA OTWARTEJ NAUKI

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?)

2. Jak zostaną uporządkowane i opisane?

   (metadane, dokumentacja)

3. 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?

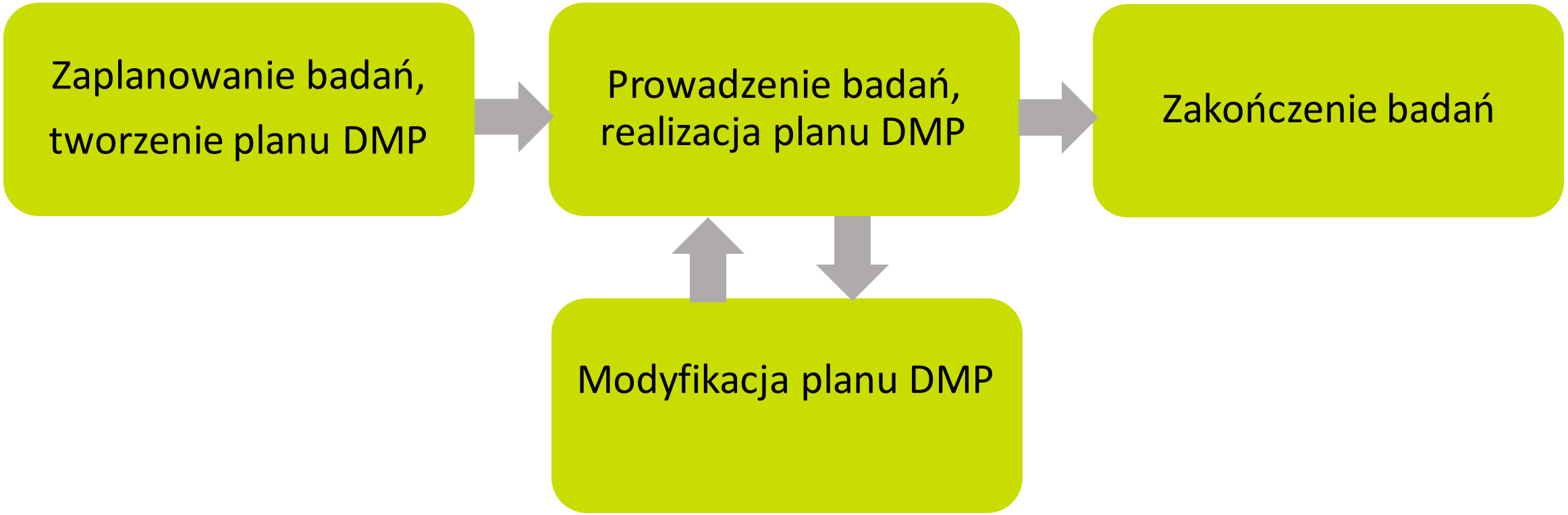
Slajd na podstawie materiałów DCC:
www.dcc.ac.uk/resources/data-management-plans/checklist

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

Slajd na podstawie: Sarah Jones, DCC, CC-BY

Zarządzanie danymi badawczymi



DataONE
Data Observation Network for Earth

https://www.dataone.org/data-management-planning

Wzorzec NSF:

1. Types of data produced
2. Data and metadata standards
3. Policies for Access and Sharing
4. Policies for Re-use, Distribution
5. Plans for Archiving and Preservation

DataONE www.dataone.org

Data Management Plan
Arthropod responses to grassland nutrient limitation.

1. Types of Data Produced
We will collect insects annually from the 30 experimental plots at each of the eight sites (see body of proposal for sampling details). Samples will be immediately deposited in sealable containers labeled with the date, site code (already existing), block, plot, and subsample. An associated record of any observations or notes will be entered in a field tablet computer and labeled with the same information. We will also record environmental information including temperature and general observations. Labeled samples will be transported back to the laboratory, where they will be sorted and identified using a dissecting microscope. We will identify and count the arthropods to the classification of order, with the exception of members of the order Auchenorrhynca, which will be identified to species or morphospecies. Identifications will be reviewed by multiple researchers associated with the project and verified with the assistance of Stuart McKamey of the Systematic Entomology Laboratory of the USDA Agricultural Research Service. Representatives of the identified species and morphospecies will be vouchered to the Bell Museum of Natural History at the University of Minnesota (U of M).

Abundance for each group will be recorded by hand in a laboratory notebook during sorting. These data will be transcribed into an Excel spreadsheet as each sample is completed. The spreadsheets will be stored on a controlled-access U of M server directory that is backed up offsite nightly. Files will be named according to the format *site_mmddyyyy_plot.csv* using existing unique site codes. Lind will be responsible for the data during and after data collection until publication.

After identification, Arthropod samples from each experimental plot will be subsampled and sent to the University of St. Thomas Kay lab for stoichiometric analysis. We will receive a spreadsheet of data after processing is complete. This spreadsheet will include the insect identification (including site code, date, year, plot, and arthropod identification) and percent by mass of carbon, phosphorus and nitrogen. These files will be saved as .csv files in the previously described server directory.

Our data set will be used in combination with the existing Nutrient Network (nutnet.unm.edu) data on plant responses to nutrient manipulation. The NutNet data is currently stored and managed in a MySQL relational database housed at the Minnesota Supercomputing Institute and accessed through a secure internet connection. We will add our data and metadata to the NutNet relational database. The existing csv files will be read into temporary tables in the MySQL database, and then inserted into permanent data tables using insert query statements. The existing database schema links tables of data observations to a "plot" table describing the experimental unit. New tables will be created for each of the arthropod data types (abundance and stoichiometry) containing the unique plot identifier. Multiple tables may be necessary for efficient data storage and management; for example, an "Arthropod" table holding scientific names for use can be used to constrain the labels of abundance records to acceptable possibilities.

2. Data and Metadata Standards
The project will leverage existing metadata standards currently stored in Ecological Metadata Language (EML) format for the NutNet project. We will add additional metadata entries for the arthropod community composition and arthropod stoichiometry; field notes taken during the time of collection will be recorded. Morpho software will be used to generate the metadata file in EML. We chose EML format for our metadata since it allows integration with existing NutNet data housed in the Knowledge Network for Biocomplexity (KNB) data repository.

3. Policies for Access and Sharing
After publication of manuscripts based on the data we collect, we will share our data and metadata with the NutNet community via data updates sent annually as .csv files from the existing central relational database. Other NutNet users will need to contact Lind for access to the data.

We will also submit both of our datasets (abundance and stoichiometry) to the U of M Digital Conservancy, an archive for digital preservation. Borer has access to this resource as a faculty member. This will occur within a year of publication. The data will be publicly available via the Digital Conservancy, which provides a permanent URL for digital documents.

4. Policies for Re-use, Distribution
Access to databases and associated software tools generated under the project will be available for educational, research and non-profit purposes. Such access will be provided using web-based applications, as appropriate.

Materials generated under the project will be disseminated in accordance with University/Participating institutional and NSF policies. Depending on such policies, materials may be transferred to others under the terms of a material transfer agreement.

Those that use the data (as opposed to any resulting manuscripts) should cite it as follows:

> Lind, E, E Borer and A Kay. yyyy. Grassland Arthropod abundance and stoichiometry associated with nutrient manipulation. [URL]; accessed on ddmmyyyy.

This information will be described in the metadata.

Intended and foreseeable users of the data are NutNet collaborators and participants, as well as other scientists interested in arthropod-plant relationships. This data set could be used in combination with similar data sets from other NutNet sites or for meta-analysis.

5. Plans for Archiving and Preservation
We will preserve both arthropod datasets generated during this project (abundance and stoichiometry) for the long term in the Digital Conservancy at the U of M. We will include the .csv files, along with the associated metadata files. We will also submit an abstract with the datasets that describe their original context and any potentially relevant project information. Borer will be responsible for preparing data for long-term preservation and for updating contact information for investigators.

icm

PLATFORMA OTWARTEJ NAUKI

DataONE
Data Observation Network for Earth

Abundance for each group will be recorded by hand in a laboratory notebook during sorting. These data will be transcribed into an Excel spreadsheet as each sample is completed. The spreadsheets will be stored on a controlled-access U of M server directory that is backed up offsite nightly. Files will be named according to the format *site_mmddyyyy_plot.csv* using existing unique site codes. Lind will be responsible for the data during and after data collection until publication.

The project will leverage existing metadata standards currently stored in Ecological Metadata Language (EML) format for the NutNet project. We will add additional metadata entries for the arthropod community composition and arthropod stoichiometry; field notes taken during the time of collection will be recorded. Morpho software will be used to generate the metadata file in EML. We

We will also submit both of our datasets (abundance and stoichiometry) to the U of M Digital Conservancy, an archive for digital preservation. Borer has access to this resource as a faculty member. This will occur within a year of publication. The data will be publicly available via the Digital Conservancy, which provides a permanent URL for digital documents.

DataONE
Data Observation Network for Earth

Access to databases and associated software tools generated under the project will be available for educational, research and non-profit purposes. Such access will be provided using web-based applications, as appropriate.

Those that use the data (as opposed to any resulting manuscripts) should cite it as follows:

> Lind, E, E Borer and A Kay. yyyy. Grassland Arthropod abundance and stoichiometry associated with nutrient manipulation. [URL]; accessed on ddmmyyyy.

This information will be described in the metadata.

icm

PLATFORMA OTWARTEJ NAUKI

Dziękuję za uwagę

Kontakt:

msommer@icm.edu.pl

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

PLATFORMA OTWARTEJ NAUKI