Open science, open data

Lidia Stępińska-Ustasiak

Open Science Platform, ICM, University of Warsaw

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

OPEN SCIENCE PLATFORM

“Open science is the idea that scientific knowledge of all kinds should be openly shared as early as is practical in the discovery process.”

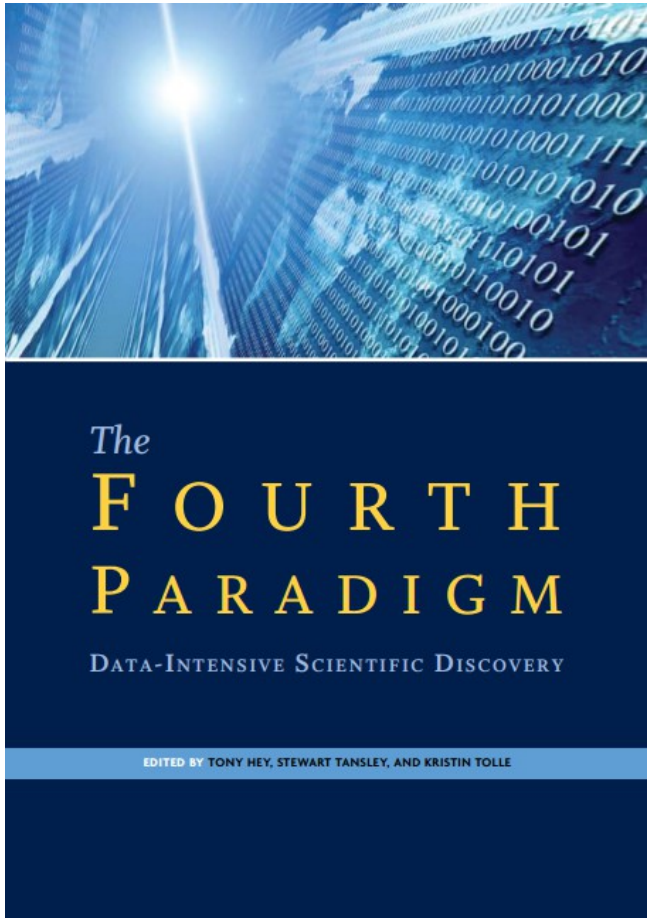
Michael Nielsen

Outline

- Openess in science
- Open research data
  - Definitions
  - Formats
  - Levels of openess
  - Depositing data
- Open Access and open research data pilot in Horizon2020
- CC licences in science
- Research Data Managment

4th Paradigm

- Empirical - describing natural phenomena (last millenia)
- Theoretical - building models and generalisations (last centuries)
- Computational - simulating complex phenomena (last decades)
- Data Exploration “data-intensive” scientific discovery (last years)

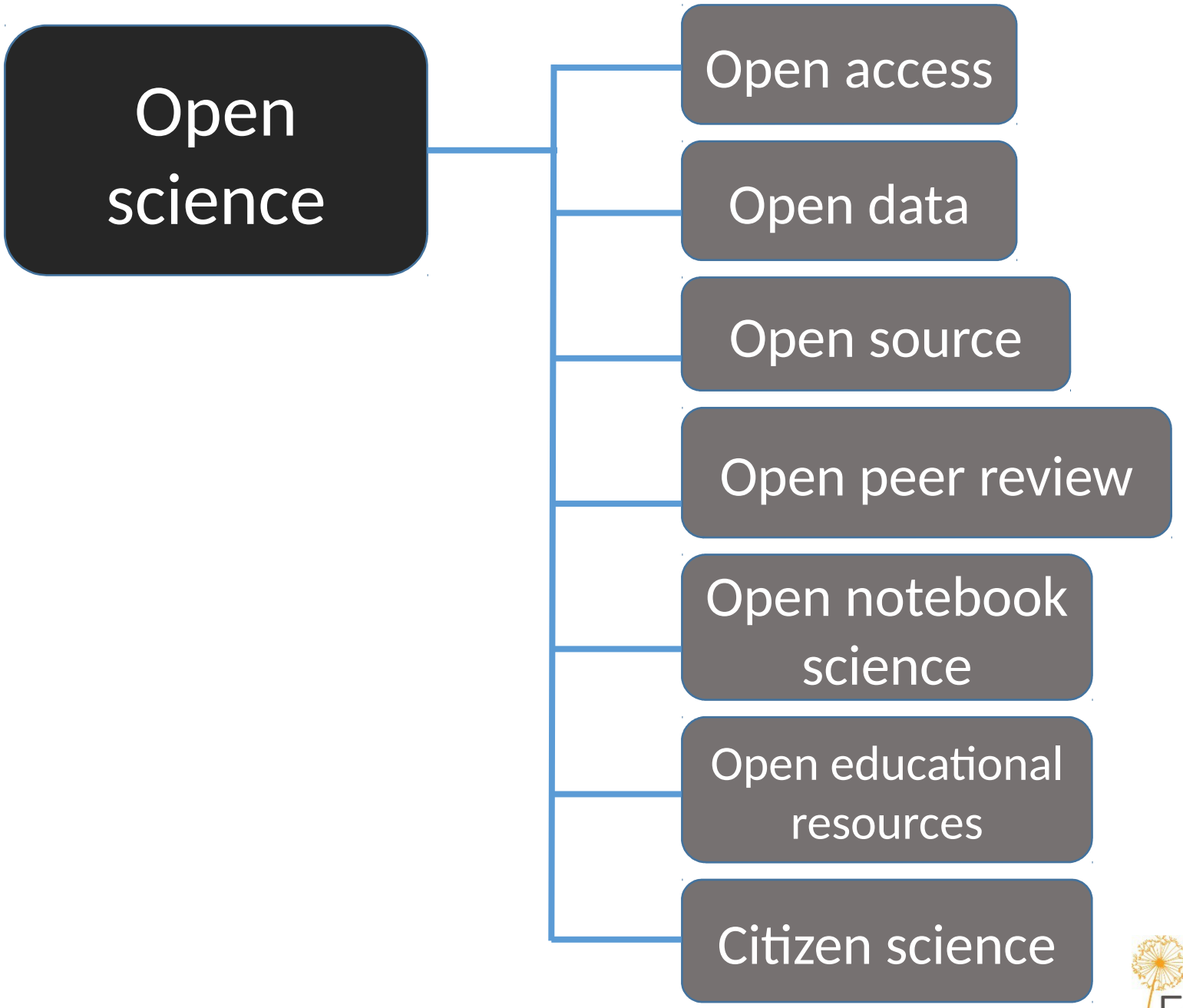


icm

FOSTER

OPEN SCIENCE PLATFORM

Scholarly communication is changing


Open science
Open access
Open data
Open source
Open peer review
Open notebook science
Open educational resources
Citizen science

What does the EC understand by the OA?

- Online access at no charge to the user
  - To peer reviewed scientific publications
  - To scientific data
- Two main publishing business models
  - Self archiving – deposit manuscripts & immediate/delayed OA provided by autho (green OA)
  - OA publishing – costs covered & immediate OA provided by publisher (gold model) e.g. „author pay” model (APC)

Objective

- The EC goal is to optimize the impact of research in Europe.

Expected benefits:

- Better and more efficient science (Science 2.0)
- Economic growth
- Broader, faster, more transparent and equal access for the benefit of researchers, industry and citizens. (Responsible Research and Innovations)

European Commission (2013):

„Open access can be defined as the practice of providing on-line access to scientific information that is free of charge to the end-user and that is re-usable.

In the context of research and innovation, 'scientific information' can refer to

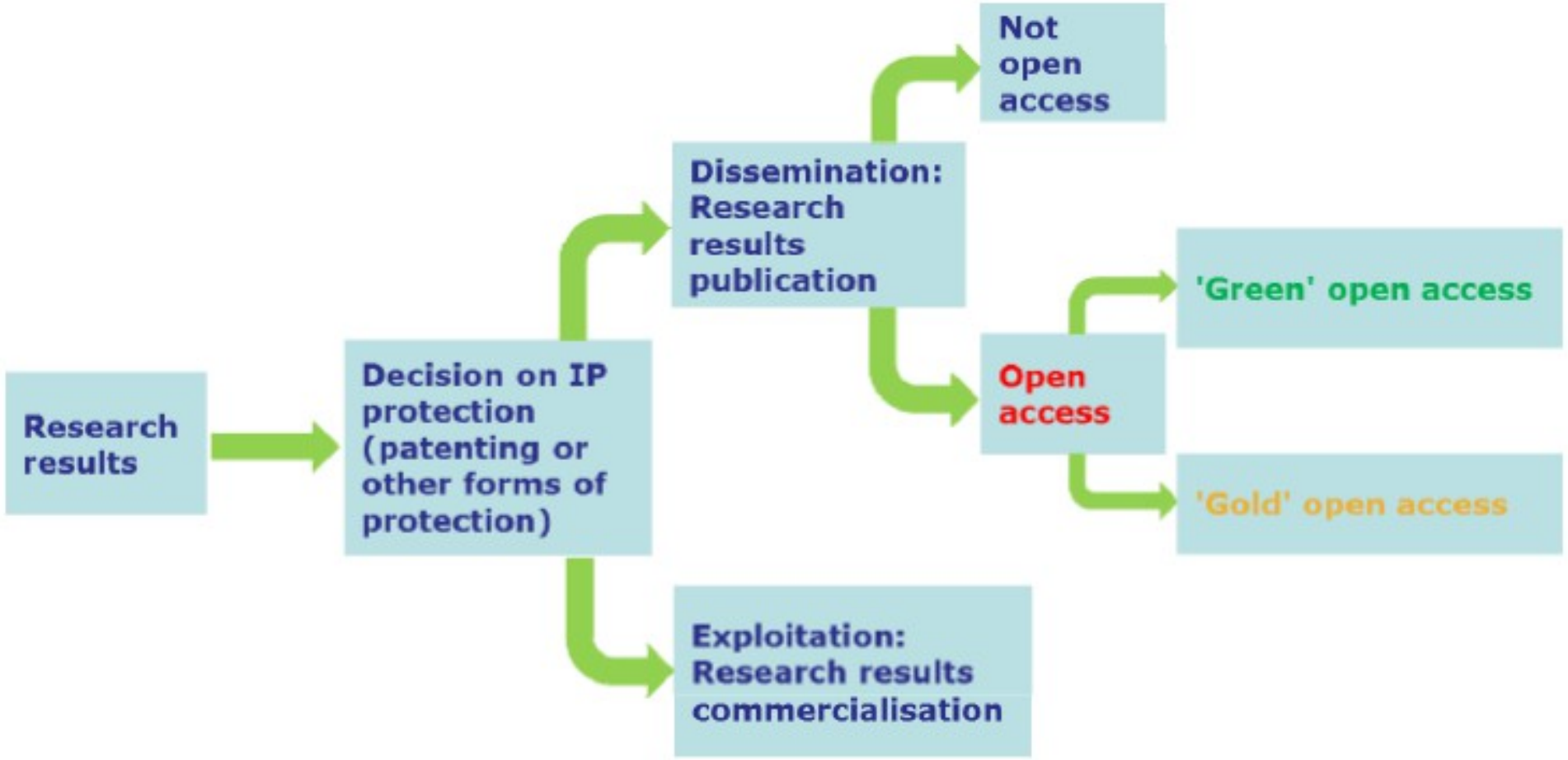
(i) **peer-reviewed scientific research articles** (published in scholarly journals)

or

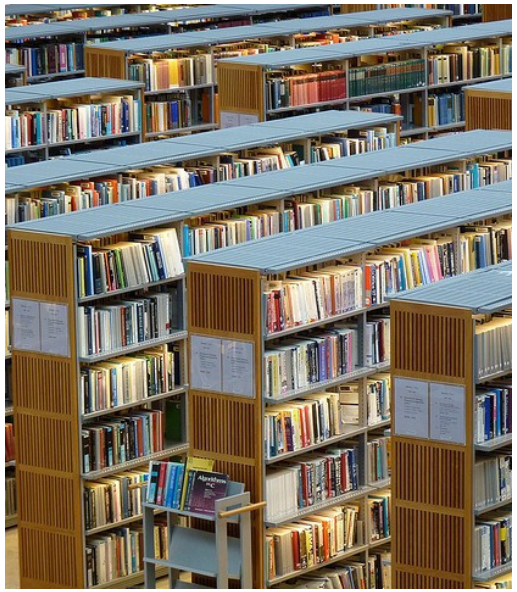
(ii) **research data** (data underlying publications, curated data and/or raw data).”

Guidelines on *Open Access* to *Scientific Publications* and *Research Data* in *Horizon 2020*. Version 16 December 2013.

Intellectual Property Rights in H2020



Scientific information

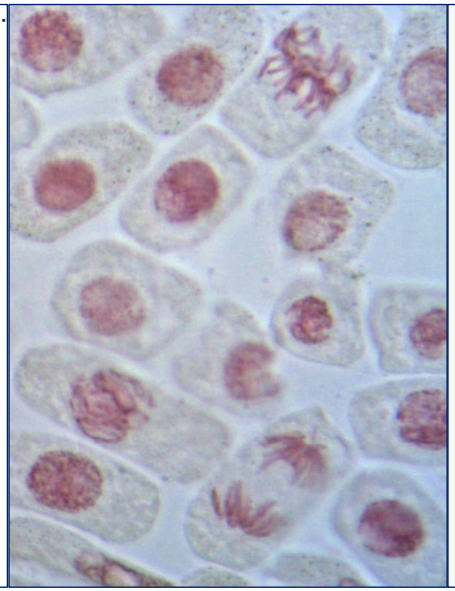


KTH Biblioteket, CC-BY-SA
https://www.flickr.com/photos/kthbiblioteket/4472640423/

Articles and books

„the recorded factual material commonly accepted in the scientific community as necessary to validate research findings”

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>
```

Research data

FOSTER

OPEN SCIENCE PLATFORM

Other definitions of research data

„...the recorded factual material commonly accepted in the scientific community as necessary to validate research findings."

„Research data is data that is collected, observed, or created, for purposes of analysis to produce original research results."

„Data is anything that has been produced or created during research."

„Anything & everything produced in the course of research”

Digital Curation Center

FOSTER

OPEN SCIENCE PLATFORM

Examples of research data:

- Numerical data
- Text documents, lab notes
- Questionnaires, responses, transcripts
- Audiotapes, videotapes
- Photographs, films
- Artefacts, specimens, samples
- Models, algorithms, scripts
- Simulation results
- Methodologies and workflows

Examples of research data

Numerical data

The focus [in the context of open access] is on research data that is available in digital form.

Models, algorithms, scripts

Simulation results

Methodologies and workflows

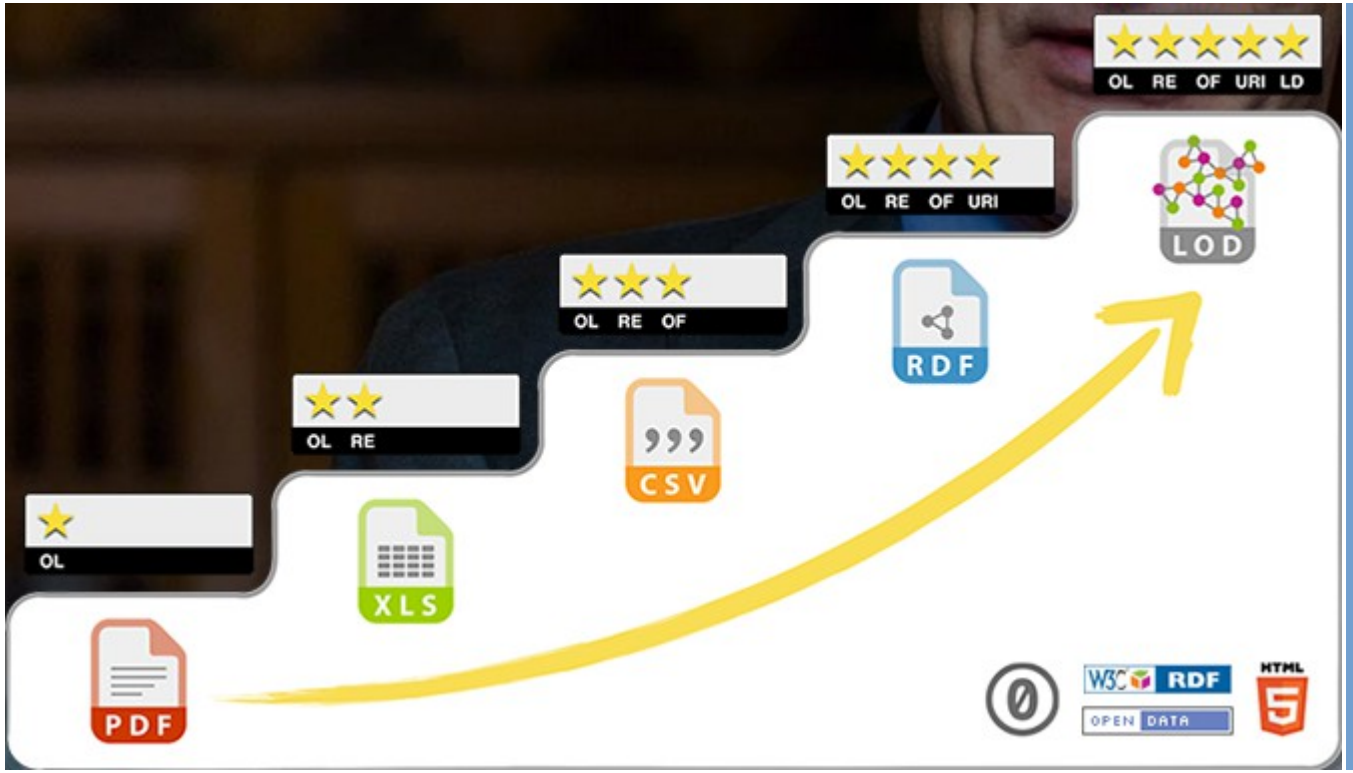
What is open data?

The Open Definition:

"Open data and content can be **freely used, modified, and shared** by **anyone** for **any purpose**."

Open Knowledge Foundation

What is open data?



★ make your stuff available on the Web (whatever format) under an open licence

★★ make it available as structured data (e.g. Excel instead of a scan of a table)

★★★ use non-proprietary formats (e.g. CSV instead of Excel)

★★★★ use URIs to denote things, so that people can point at your stuff

★★★★★ link your data to other data to provide context

Tim Berners-Lee, 5-star Open Data, 5stardata.info

→ This model is concerned with removing technical barriers to data re-use.

Formats

Type of data	Reccomended	Avoid for data sharing
Tabular	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	Qiucktime
Images	TIFF, JPEG2000, PNG	Gif, JPG
Structured data	XML, RDF	RBDMS

Major sources of open data

Public data

Statistical data

Financial

Cultural

Climate

Environment

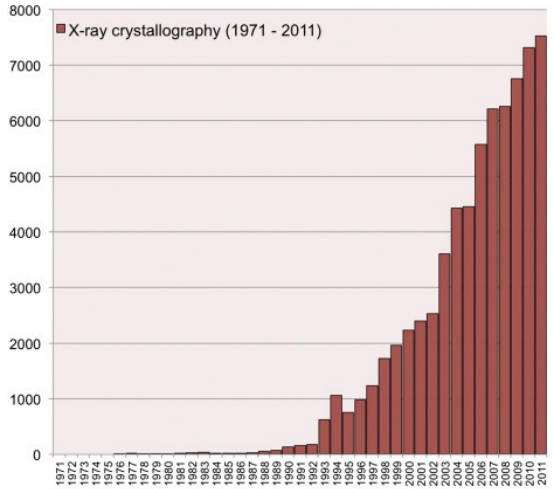
Transport

…

Research data

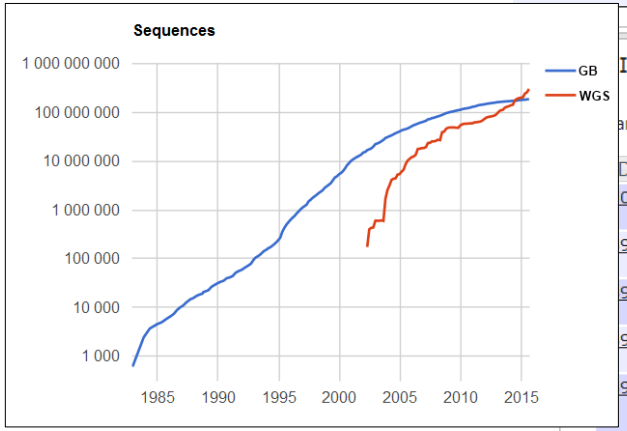
Specialized data repositories

Protein Data Bank – since 1971



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

GenBank – since 1982



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

Oxford Text Archive – since 1976

What about data for which no specialized repositories exist?

→ Broad or general data repositories

DRYAD

Data Archiving and Networked Services DANS EASY

ceON RepOD REPOZYTORIUM OTWARTYCH DANYCH

Purdue University Research Repository

zenodo

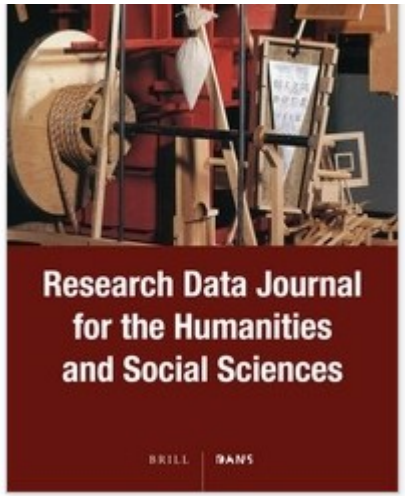
figshare

→ Data journals

SCIENTIFIC DATA

npg nature publishing group

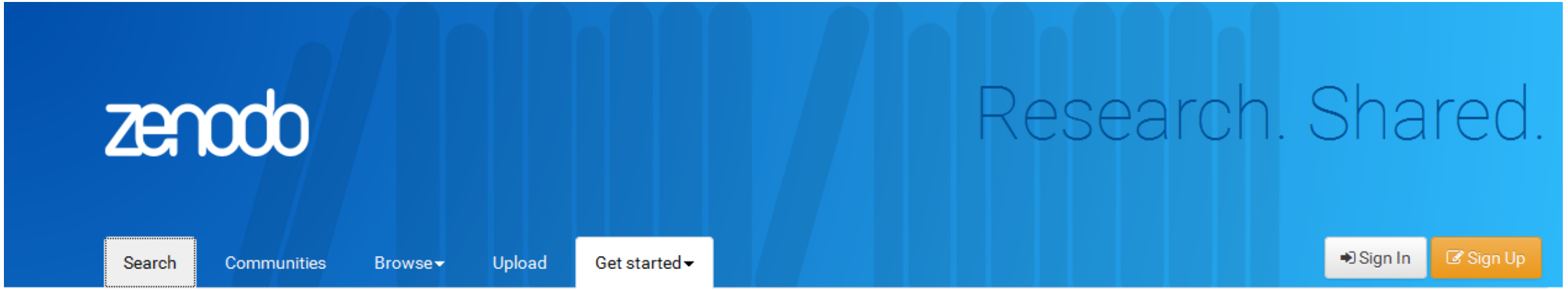
ELSEVIER Data in Brief



icm

FOSTER

OPEN SCIENCE PLATFORM



Home / About

About Zenodo

Zenodo builds and operates a simple and innovative service that enables researchers, scientists, EU projects and institutions to share, preserve and showcase multidisciplinary research results (data and publications) that are not part of the existing institutional or subject-based repositories of the research communities.

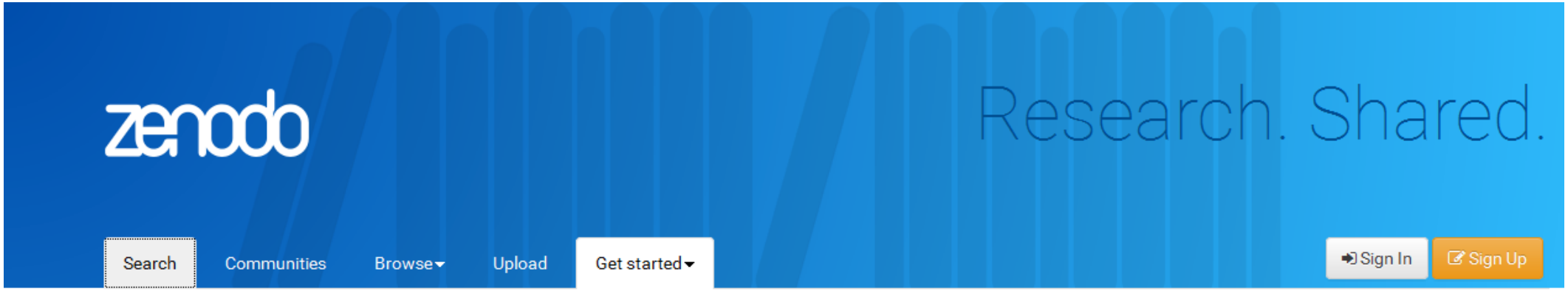
Zenodo enables researchers, scientists, EU projects and institutions to:

- easily share the long tail of small research results in a wide variety of formats including text, spreadsheets, audio, video, and images across all fields of science.
- display their research results and receive credit by making the research results citable and integrating them into existing reporting lines to funding agencies like the European Commission.
- easily access and reuse shared research results.

The name

Zenodo is derived from Zenodotus, the first librarian of the Ancient Library of Alexandria and father of the first recorded use of metadata, a landmark in library history.

Logo



Zenodo enables researchers, scientists, EU projects and institutions to:

- easily share the long tail of small research results in a wide variety of formats including text, spreadsheets, audio, video, and images across all fields of science.
- display their research results and receive credit by making the research results citable and integrating them into existing reporting lines to funding agencies like the European Commission.
- easily access and reuse shared research results.

icm

FOSTER

OPEN SCIENCE PLATFORM

ZENODO

- Zenodo is a free-to-use data archive, run by CERN
- It accepts any kind of data, from any academic discipline
- It is generally preferable to store data in a disciplinary data centre, but not all scholarly subjects are equally well served with data centres, so this may make for a useful fallback option
- See http://zenodo.org/ for more details

Should all data be open?

Should all data be open?
No

Privacy protection (human subjects!)

National security issues

Protection of endangered species, of archaeological sites, etc.

Interference with commercialization plans

But data existence should always be open:

- Allows discovery & negotiation on use
- Avoids pointless replication

https://www.youtube.com/watch?v=RGtPVIBmFBI&feature=youtu.be

Why data sharing is worth your attention?

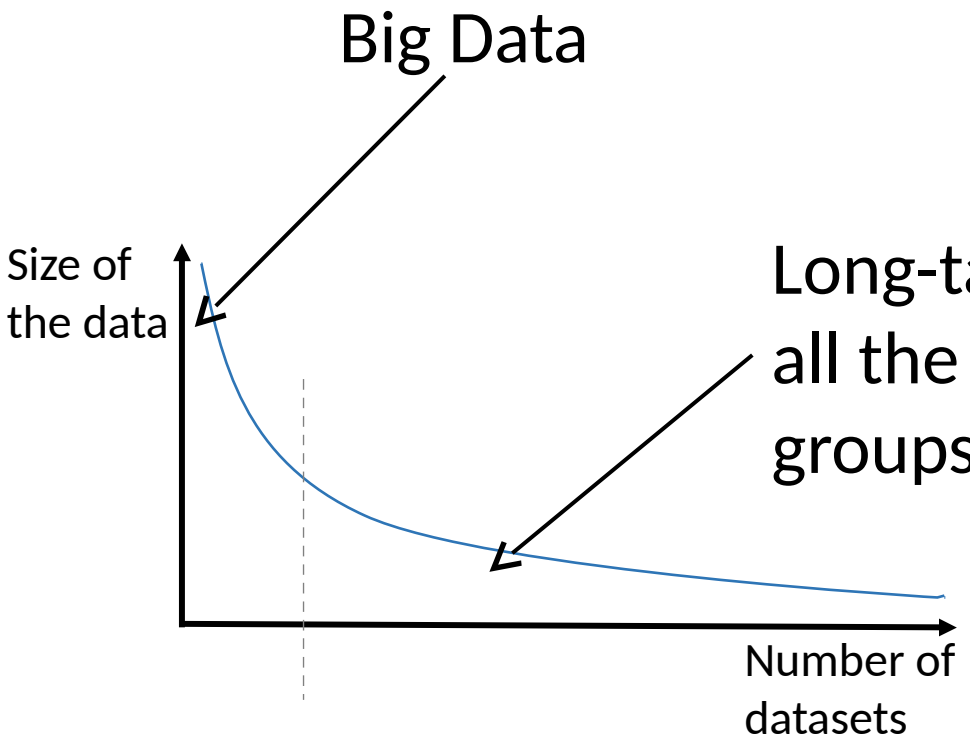
- Digital technology now used very widely in research, and is enabling new research and scientific paradigms
- Research funders and publishers know that digital research data can be expensive to produce but inexpensive to share, making reuse more feasible and desirable
- The challenge is to ensure digital research findings can be reproduced and cited

icm

FOSTER

OPEN SCIENCE PLATFORM

The long tail of research data



Long-tail of data:
all the data produced by small research groups and individual researchers

„To me, the really difficult challenge is (…) the variety. The heterogeneity, as you put it. And we see this particularly in what they call the long tail of data (…)”

Mark Parsons, Research Data Alliance

Excercise

Objections to data sharing

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

OPEN SCIENCE PLATFORM

How to answer to the most commonly heard objections to data sharing?

1. My data in not of interest or use to anyone else.

Replies (1)

- It is! Researchers want to access data from all kinds of studies, methodologies and disciplines. It is very difficult to predict which data may be important for future research. Your data! May also be essential for teaching purposes. Sharing is not just about archiving your data but about sharing them amongst colleagues.

2. I want to publish my work before anyone else sees my data.

Replies (2)

- Data sharing will not stand in the way of you first using your data for your publications. Most research funders allow you some period of sole use, but also want timely sharing. Also remember that you have already been working with your data for some time so you undoubtedly know the data better than anyone coming to use them afresh. If you are still concerned you can embargo your data for a specific period of time.

3. If I ask my respondents for consent to share their data, then they will not agree to participate in the study.

Replies (3)

- Don’t assume, that participants will not participate because data sharing is discussed. Talk to them, they may be less reluctant than you might think or less concerned over data sharing. Make it clear that is entirely their decision. Explain that data sharing means and why it might be important.
- If you not have asked for permission during research you can return to gain retrospective permission from participants.

4. I’m doing quantitative research and the combination of my variables discloses my participants’ identities.

Replies (4)

- Quantitative data can by anonymised trough processes of aggregation, top coding, removal of variables or controlled access to certain variables.

5. I have collected audio-visual data and I cannot anonymise them, therefore I cannot share these data.

Replies (5)

- Visual data can be anonymised trough blurring faces or distorting voices but it can be time consuming. It can mean losing much of the value of the data. It is better to ask for consent to share data from participants to share data in unanonymised form or / and control access to the data.

6. I’m doing highly sensitive research. I cannot possibly make my data available for others to see.

Replies (6)

- Ask respondents and see if you can get consent for sharing in the first instance. Anonymisation procedures can help to protect identifying information. If this two tactics are not apropriate. Than consider controlling access to tha data or embargoing for a period of time.

7. It is impossible to anonymise my transcripts as too much information is lost.

Replies (7)

- Sometimes access control on the data may be a better solution than anonymisation if too much useful information would be lost.

8. My data collection contains the data which I have purchesed and it cannot be made public.

Replies (8)

- It is important to know who holds the copyright to the data you are using and to obtain relevant permissions. You need to be aware of the licence conditions of the data you are using and what you can and cannot do with the data.

9. Other researchers would not understand my data at all or may use them for a wrong purpose.

Replies (9)

- Producing good documentation and providing contextual information for your research project should enable other researchers to corretly use and understand your data.

10. There is IPR in the data.

Replies (10)

- This should not be a problem if you seek copyright permission from the owner of the intellectual property rights. This is best done early on in the research project but also may be done retrospectively.

Role playing exercise derived from the UKDA's "Potential barriers to data sharing – with suggested solutions" (CC-BY-NC-SA) The original is available from http://data-archive.ac.uk/create-manage/training- resources

Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020

Version 1.0
11 December 2013

Open Access in Horizon 2020

Mandate on open access to publications:

„Under Horizon 2020, each beneficiary must ensure open access to all peer-reviewed scientific publications relating to its results.”

Open Access in Horizon 2020

In order to comply with this requirement, beneficiaries must, **at the very least**, ensure that their publications, if any, can **be read online, downloaded and printed**. However, as any **additional rights** such as the right to **copy, distribute, search, link, crawl, and mine** increase the utility of the accessible publication, beneficiaries should **make every effort** to provide for as many of them as possible.

icm

FOSTER

OPEN SCIENCE PLATFORM

Open Access in Horizon 2020

Open research data pilot:

„The Open Research Data Pilot applies to two types of data:

1) the **data** (…) **needed to validate the results presented in scientific publications** as soon as possible;

2) **other data** (…) as specified and within the deadlines laid down in the data management plan."

„Participating projects are required to deposit the research data described above, preferably into a research data repository."

Open Access in Horizon 2020

Open research data pilot:

- Only for projects from 7 selected areas.
- You can opt-in, and you can also opt-out.

described above, preferably into a research data repository.”

Open Access in Horizon 2020

Participating projects are required to deposit the research data described above, preferably into a research data repository.

As far as possible, projects must then take measures to enable for third parties to access, mine, exploit, reproduce and disseminate (free of charge for any user) this research data.

One straightforward and effective way of doing this is to attach a Creative Commons Licence (CC-BY or CC0 tool) to the data deposited.

H2020 - areas participating in the data pilot

- Future and Emerging Technologies
- Research infrastructures – part e-Infrastructures
- Leadership in enabling and industrial technologies – Information and Communication Technologies
- Societal Challenge: 'Secure, Clean and Efficient Energy' – part Smart cities and communities
- Societal Challenge: 'Climate Action, Environment, Resource Efficiency and Raw materials' – except raw materials
- Societal Challenge: 'Europe in a changing world – inclusive, innovative and reflective Societies'
- Science with and for Society Projects in other areas can participate on a voluntary basis

icm

FOSTER

OPEN SCIENCE PLATFORM

Reasons for opting out

- If results are expected to be commercially or industrially exploited
- If participation is incompatible with the need for confidentiality in connection with security issues
- If incompatible with existing rules on the protection of personal data
- Would jeopardise the achievement of the main aim of the action
- If the project will not generate / collect any research data
- If there are other legitimate reasons to not take part in the Pilot

Can opt out at proposal stage OR during lifetime of project.

Should describe issues in the project Data Management Plan.

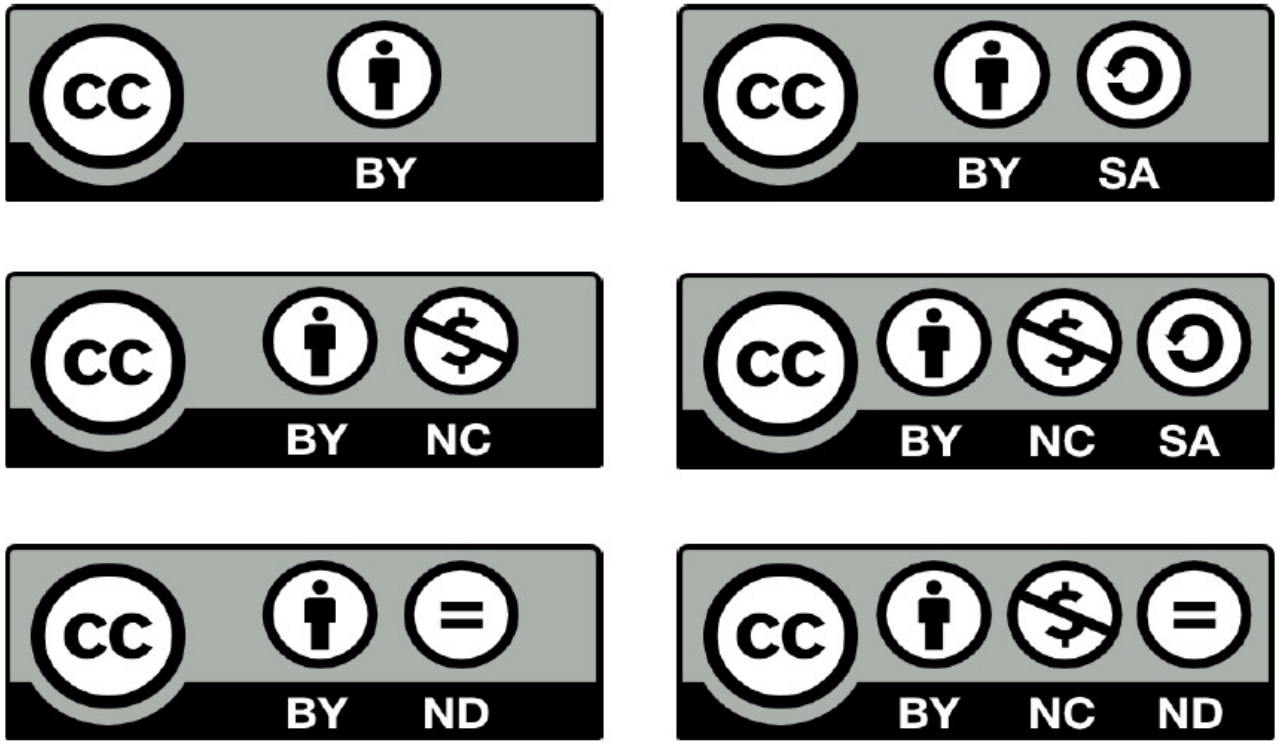
Legal aspects

CC licences

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

OPEN SCIENCE PLATFORM

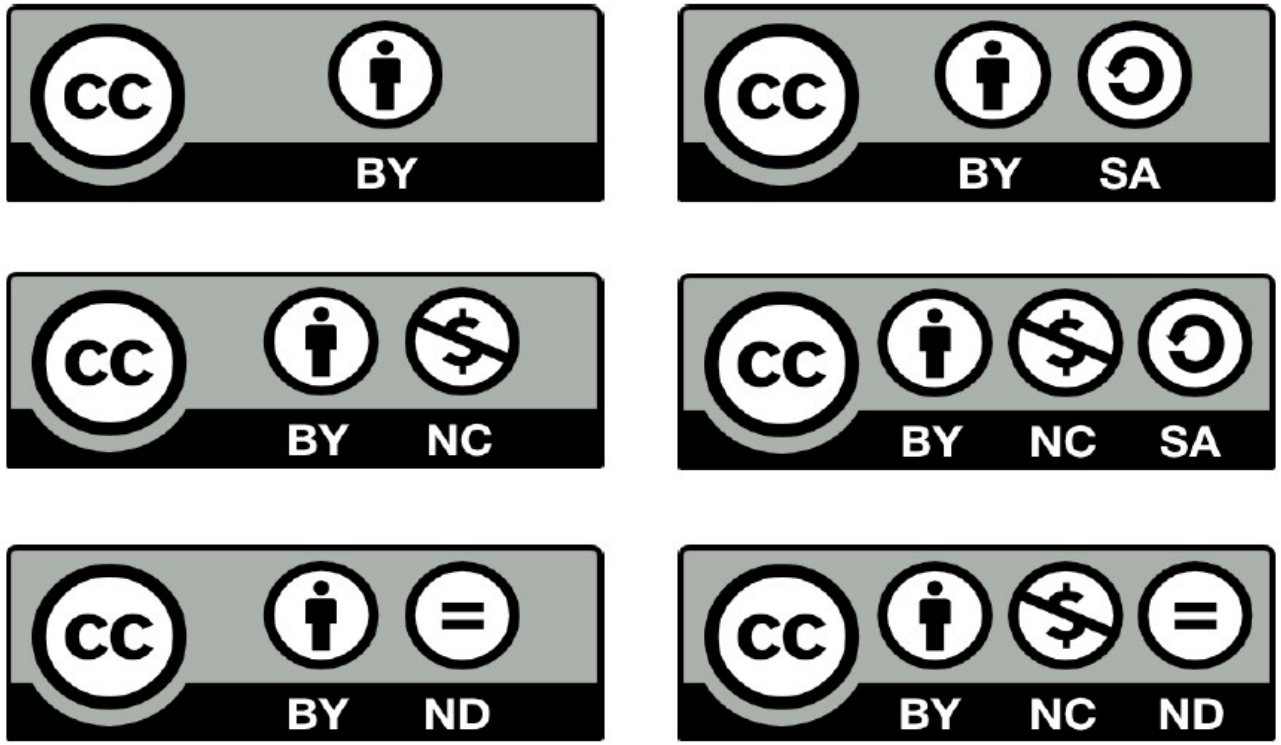
What are Creative Commons Licenses?



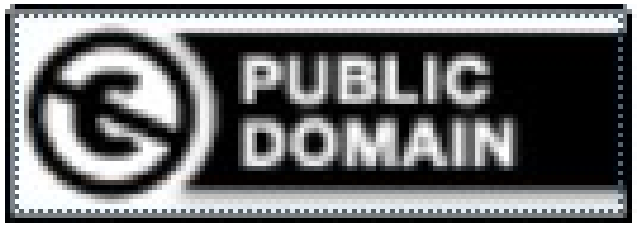
What are Creative Commons Licenses?

BY – Attribution **NC – Non-commercial**

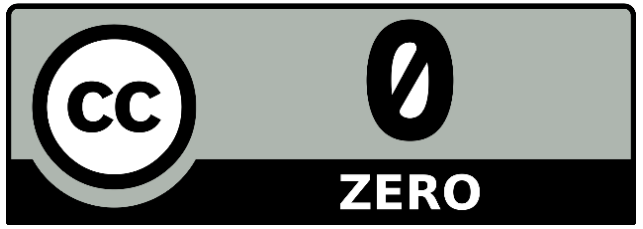
SA – Share Alike **ND – No derivatives**



Public Domain



Public Domain Mark



Public Domain Dedication

Gratis

open access

the right to read

Libre

open access

the right to read
and re-use

CC0 is easy to use

You don’t need to know what rights actually apply to your dataset

(what is protected?)

 you should know this for CC-BY (and other CC licenses)

Why CC0 for research data?

BY: Datasets are particularly prone to attribution stacking, where a derivative work must acknowledge all contributors to each work from which it is derived, no matter how distantly.

SA: The problem with copyleft licences is they prevent the licensed data being combined with data released under a different copyleft licence: the derived dataset would not be able to satisfy both sets of licence terms simultaneously.

NC: Non-commercial licences may have wider implications than intended due to the ambiguity of what constitutes a commercial use.

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

Open Access in Horizon 2020

Open research data pilot:

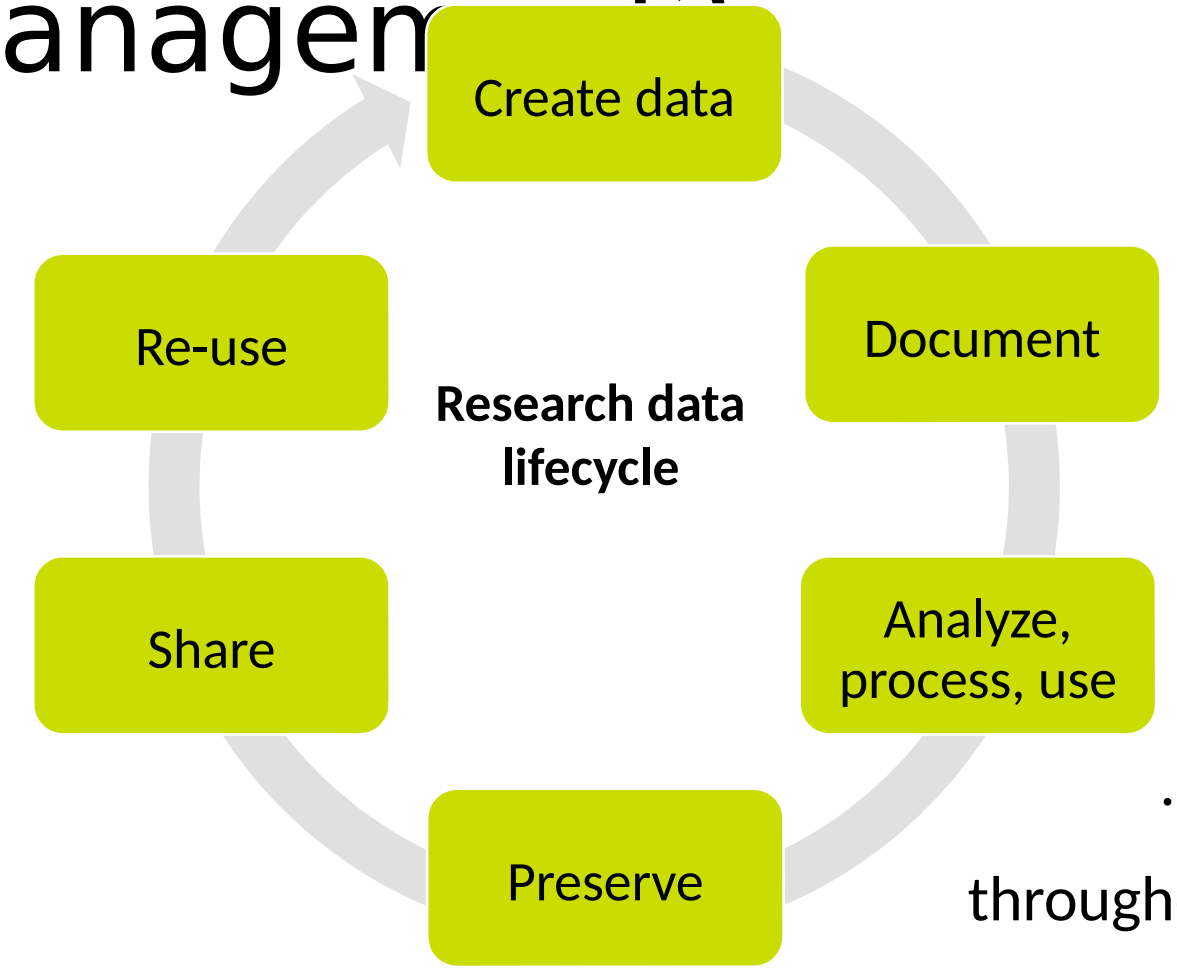
„The use of a detailed data management plan covering individual datasets is required for funded projects participating in the Open Research Data Pilot.”

Research data management

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

OPEN SCIENCE PLATFORM

What is Research Data Management?



...an active approach towards handling data throughout all stages of the research data lifecycle.

Active data management

- Data management planning
- Creating data
- Documenting data
- Accessing & using data
- Storage and backup
- Selecting what to keep
- Sharing data
- Data licencing and citation
- Preserving data
- …

Digital Curation Center

Data Selection – guidelines

1. ***Legal requirements*** to retain the data beyond its immediate use.
2. ***Scientific or Historical Value***: this involves inferring anticipated future use.
3. ***Uniqueness***: does it duplicate existing datasets?
4. ***Non-Replicability***: would it be feasible to replicate the data? (high costs, one-time events)
5. ***Potential for Redistribution***: the reliability, integrity, and usability of the data files (do formats meet technical criteria? are IPRs addressed?)
6. ***Economic Case***: costs for managing and preserving the data are justifiable when assessed against evidence of potential future benefits.
7. ***Full documentation***: documentation is comprehensive and correct.

Based on:
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

File formats - tactic

If you want your data to be re-used and sustainable in the long-term, you typically want to opt for open, non-proprietary formats.

- Do you have a choice or do the instruments you use only export in certain formats?
- What is common in your field? Try to use something that is accepted and widespread
- Does your data centre recommend formats? If so it's best to use these.

Data selection...

...depends on what researchers want to do with their data;

what they are allowed to do with the data;

and what the institution can afford to do with the data.

What is a DMP?

A brief plan that outlines

- what data will be created and how
- how it will be managed (storage, back-up, access…)
- plans for data sharing and preservation

Lots of research funders require DMP

Why develop a DMP?

DMPs are useful whenever researchers are creating data to:

- Make informed decisions to anticipate and avoid problems
- Avoid duplication, data loss and security breaches
- Develop procedures early on for consistency
- Ensure data are accurate, complete, reliable and secure
- Save time and effort

Five common themes

1. Description of data to be collected / created
   (i.e. how will it be collected, content, type, format, volume…)

2. Documentation & metadata
(standards and formats, structure of file naming, etc.)

3. Ethics and Intellectual Property
   (highlight any restrictions on data sharing e.g. privacy, confidentiality)

4. Plans for data sharing and access
   (i.e. how, when, to whom)

5. Strategy for long-term preservation

www.dcc.ac.uk/resources/data-management-plans/checklist

Advice on writing DMPs

- Keep it short and simple, but be specific
- Seek advice - consult and collaborate
- Base plans on available skills and support
- Make sure implementation is feasible
- **Remember**: plans change and should evolve

For better understanding of your data

- Think about what is needed in order to find, evaluate, understand, and reuse the data.
- Have you documented what you did and how?
- Did you develop code to run analyses? If so, this should be kept and shared too.
- Is it clear what each bit of your dataset means? Make sure the units are labelled and abbreviations explained.
- Record metadata so others can find your work e.g. title, date, creator(s), subject, format, rights…,

Which data need to be kept

- Could this data be re-used
- Must it be kept as evidence or for legal reasons
- Should it be kept for its potential value
- Consider costs – do benefits outweigh cost?
- Evaluate criteria to decide what to keep

- 5 steps to decide what data to keep www.dcc.ac.uk/resources/how-guides/five-steps-decide-what-data-keep

Where to deposit?

- Does your publisher or funder suggest a repository?
- Are there data centres or community databases for your discipline?
- Does your university offer support for long-term preservation?

In this section

- Briefing Papers
- How-to Guides & Checklists
- Developing RDM Services
- Curation Lifecycle Model
- Curation Reference Manual
- Policy and legal
- **Data Management Plans**
  - Checklist
  - DMPonline
  - FAQ on DMPonline
  - FAQ on Data Management Plans
  - Funders' requirements
  - Guidance and examples
- Tools
- Case studies
- Repository audit and assessment
- Standards
- Publications and presentations

DMPonline

Research funders and organisations increasingly require data management plans, both during the bid-preparation stage and after funding has been secured.

DMPonline is the DCC's data management planning tool. It provides tailored guidance and examples to help researchers write data management plans.

The tool includes a number of templates for funders in the UK and overseas so researchers can write DMPs according to the specific requirements they need to meet. It can also be customised by institutions so they can add their own templates and guidance.

A screencast provides an overview of how the tool works.

Try the tool for yourself at http://dmponline.dcc.ac.uk

Anyone can use DMPonline. If your organisation is not listed, just select 'other organisation' or ask for it to be added.

If you would like to create a foreign language version of DMPonline, please contact us on dmponline@dcc.ac.uk

Useful links

DMPonline

IDCC

The 10th International Digital Curation Conference (IDCC) took place at 30 Euston Square in London, UK, on 9 - 12 February 2015.

This year's theme was "Ten years back, ten years forward: achievements, lessons and the future for digital curation"

We'll soon be announcing where IDCC16 will be located!

Read more

Excercise

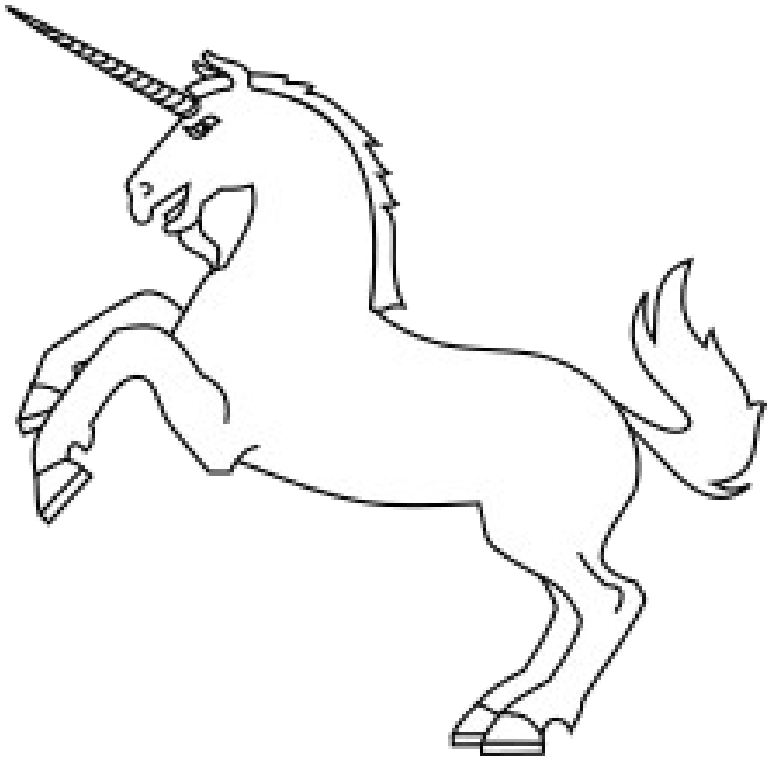
Define and select your data

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

Choose one specific research project and for this project:

1. Define what data will be generated (all of it!)
2. What would you select for preservation?
3. How would you share your data?

There is no such thing as ideal data.



icm

FOSTER

OPEN SCIENCE PLATFORM

Thank you for your attention

Contact:

l.stepinska-ustasiak@icm.edu.pl

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

OPEN SCIENCE PLATFORM