Ensuring compliance with the FAIR principles in access to research data by building services for data management in scientific institutions

Dr Anna Wałek
Gdańsk University of Technology Library

ALLDATA 2022 --- Barcelona 24-28.04.2022
Open Science
Open Applications
Open Notebooks
Open Source
Open Methods
Open Data
Open Peer Review
Open Manuscripts
Open Educational Resources
Open Access
Open citations
Open scientific social networks
Open assessment
Citizen Science
Research results publication

Research

- Dissemination plan
- Research results
- Data management plan

Decision to disseminate / share

- Publications
  - Gold Open Access
  - Green Open Access

- Depositing research data
  - Access and use free of charge
  - Restricted access and/or use

Decision to exploit / protect

- Patenting (or other form of protection)
Plan S

Accelerating the transition to full and immediate Open Access to scientific publications
Open Science

An approach to the scientific process that focuses on spreading knowledge as soon as it is available using digital and collaborative technology. Expert groups, publications, news and events.

The EU's open science policy

Open science is a policy priority for the European Commission and the standard method of working under its research and innovation funding programmes as it improves the quality, efficiency and responsiveness of research.

When researchers share knowledge and data as early as possible in the research process with all relevant actors it helps diffuse the latest knowledge.

And when partners from across academia, industry, public authorities and citizen groups are invited to participate in the research and innovation process, creativity and trust in science increases.

That is why the Commission requires beneficiaries of research and Innovation funding to make their publications available in open access and make their data as open as possible and as closed as necessary. It recognises and rewards the participation of citizens and end users.
“Data is a renewable resource as much as sun and wind. Every 18 months we double the amount of data we produce, 85 per cent of which is never used.”
Polish National Science Centre

One of the signatories of PlanS

DMP’s in the project proposals since 2020
What is research data:

Research data are the original sources or material that you have created or collected to conduct your research project.

They can be digital or non-digital.

The response to your research question is based on the analysis of these research data.
types of data:

- qualitative (dealing with things numerically)
- quantitative (descriptive in nature and dealing with the quality, categorization).

- primary data (original data that arise from a particular experiment or observation), which are gathered and maintained by researchers
- secondary data, often used by researchers but originally created by someone else

- observational data – data that have been gathered from observing a particular phenomenon,
- experimental data – in contrast – derived from controlled, randomized experiments.
WHAT IS THAT?

DATA FOR FUTURE GENERATIONS

NO IDEA

DATA?

PUBLICATIONS AND DATA

1st

ARCHAEOLOGY

(?!)

OVER THERE

2

3
The Data Citation Index™ provides a single point of access to quality research data from global repositories across disciplines. Descriptive records are created for data objects and linked to literature articles in the Web of Science.™

As data citation practices increase, the Data Citation Index aims to provide a clearer picture of the full impact of research output, as well as to act as a significant tool for data attribution and discovery.
Research Data Lifecycle
(created by the UK Data Archive)

It covers the lifespan of research data from the moment of its creation through the reuse of the data.
Researchers perspective

- Creating data
- Processing data
- Analyzing data
- Preserving data
- Reusing data
- Giving access to data
Curator’s perspective

Curation lifecycle model
(created by The Digital Curation Centre)

all the processes and components involved in data curation from an archivist’s or a curator’s perspective
Curator’s perspective

1. creating/receiving data
2. selecting data
3. ingesting
4. storing data
5. performing preservation actions
6. accessing data for use and reuse
7. transforming data
Data management plan

Data Management Plan (DMP)

show how research data will be collected, organized, managed, and preserved during the project and after.
A DMP is.....?

.....a written document describing how DATA of a research project is managed during the life-cycle.
Legal issues

### Problems Occurred During Preparing DMPs in Different Disciplines

<table>
<thead>
<tr>
<th>Biology</th>
<th>Civil Engineering</th>
<th>Computer Science</th>
<th>Chemistry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethical issues e.g. animal testing</td>
<td>Fear of being scoped</td>
<td>Eligibility to hold rights to a database</td>
<td>Ethical issues with collecting research data e.g. clinical trials, patient privacy</td>
</tr>
<tr>
<td>Replication of research</td>
<td>Not knowing where to share research and technical data</td>
<td>Different practices of sharing code e.g. via GITHUB</td>
<td>Concerns about data mining</td>
</tr>
<tr>
<td>Fear of misuse</td>
<td>Large amount of data is collected by instrumentation — authorships' concerns</td>
<td>Data versioning</td>
<td>Secondary analysis of data</td>
</tr>
<tr>
<td>Not enough capacity in repositories to share data</td>
<td>Construction data is rather complex and multiple engineering parties are often involved</td>
<td>Hard to share ever-greater quantities of data</td>
<td>Desire to protect confidential commercial information</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Economics and Management
- Contractual obligations
- GDPR
- Data value cycle is very complex and might involve numerous stakeholders and different business agreements
- Complications with quality data anonymization e.g. interviews
### Legal arrays

<table>
<thead>
<tr>
<th>Category</th>
<th>Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patent law</td>
<td>• What has to be considered when research data (can) reach patentability?</td>
</tr>
<tr>
<td>Science law</td>
<td>• Can research data be licensed and published on a mandate basis?</td>
</tr>
<tr>
<td>Contracts</td>
<td>• Are there any “intellectual property” agreements with research data?</td>
</tr>
<tr>
<td>Copyright protection</td>
<td>• Do research data fall under copyright law at all?</td>
</tr>
<tr>
<td>Constitutional rights</td>
<td>• What constitutional limits must be observed?</td>
</tr>
<tr>
<td>International law</td>
<td>• Which legal regulations exist outside Poland?</td>
</tr>
<tr>
<td>Conditions of the funders</td>
<td>• What are the conditions imposed by sponsors (NCN, industry)?</td>
</tr>
<tr>
<td>Privacy protection</td>
<td>• What research data is worth protecting?</td>
</tr>
<tr>
<td>EU law</td>
<td>• What does the European data economy contribute for research data?</td>
</tr>
<tr>
<td>Policies</td>
<td>• Which legal obligations can policies develop?</td>
</tr>
</tbody>
</table>

Based on Katarzyna Biernacka 11.04.2019
DATA FLOW

Generate and prepare data files

Deposit dataset to repository

Data checked and curated

Published dataset

AUTHOR MANAGED FUNCTIONS

Write experimental article

Submit manuscript to publisher

ARTICLE FLOW

Author includes data citation with article submission

Article reviewers access data on repository

Discoverable links via DOIs

Article and data reviewed

Published article
Opendatahandbook.org defines open data as:

*data that can be freely used, re-used and redistributed by anyone – subject only, at most, to the requirement to attribute and sharealike.*

Opendefinition.org says that:

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

As matter of principle, adhere to the habit of making your **as open as possible, as closed as necessary.**
Open Data

Can be freely used, modified & shared by anyone for any purpose

http://opendefinition.org

Limits on who can access & use data, how, or for what purpose
- only certain (types of) users
- only certain types of use
- ...

Under embargo
Unable to share
FAIR Data

Findable, Accessible, Interoperable, Reusable
FAIR Data

Findable
Metadata and data should be findable for both humans and computers

Interoperable
Data needs to work with applications or workflows for analysis, storage and processing

Accessible
Once found, users need to know how the data can be accessed

Reusable
The goal of FAIR is to optimise data reuse via comprehensive well-described metadata
The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
   A1.1. the protocol is open, free and universally implementable
   A1.2. the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared and broadly applicable language for knowledge representation
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
   R1.1. (meta)data are released with a clear and accessible data usage licence
   R1.2. (meta)data are associated with detailed provenance
   R1.3. (meta)data meet domain-relevant community standards
Findable
Persistent Identifiers (PIDs)
Rich metadata
Indexed data repositories
PID s in metadata

Accessible
Standard communications protocol
Open, free protocol
Authentication, where necessary
Metadata is always available

Interoperable
Vocabularies
Vocabularies are FAIR
Linked metadata

Reusable
Metadata have multiple attributes
Usage license
Provenance
Community standards
Findable Accessible Interoperable Reusable ≠ Open
How do Open and FAIR intersect?

- Open
- FAIR data
- Managed data
- Internal Self-interest
- External Community benefit

Core Certified Repositories

Applications are made public only once certification of a data repository has been approved by the CoreTrustSeal Board. Certification is against the version of the Core Trustworthy Data Repositories Requirements named in the link to the public application (e.g., 2017-2019). The CoreTrustSeal for Data Repositories is valid for three years from the certification date listed within the public application.
Open Research Data Repository

The research data repository provides additional information on its service.

The research data repository was reviewed by the re3data.org Team.

The research data repository is either certified or supports a repository standard.

The research data repository provides a policy.

The research data repository provides open/restricted/closed access to its data.

The terms of use and licenses of the data are provided by the research data repository.

The research data repository uses a persistent identifier system to make its provided data persistent, unique and citable.
academic libraries serve the academic community as important partners in research process and information management
“Research data services are services that address the full data lifecycle, including the data management plan, digital curation (selection, preservation, maintenance, and archiving), and metadata creation and conversion” (Tenopir, Sandusky, Allard and Birch, 2012).
Data librarian – skills

- high level communication
- advocacy
- negotiation and capability building skills
Data librarian – skills

- comfortable with change,
- have a service philosophy,
- willingness to learn,
- discretion,
- “boundless curiosity”,
- be adaptable,
- assertive
Data librarian – skills

contextual knowledge

- research environment
- related funding agency policies
- research measures
- research evaluation activities
Data librarian – skills

Being open to:

- Discipline specific research life cycles
- Cultures, processes,
- Ethics,
- Disciplinary research methods,
- Intellectual property, licensing and policies,
- Access norms and cultural sensitivities
Data librarian – skills

- facilitate data sharing,
- linked data,
- data management lifecycle,
- quality control,
- data processing,
- data management planning
- ability to understand and support data storage
Data librarian – skills

Skills in:
- Applying metadata,
- Metadata standards,
- Ontologies,
- Structured information,
- Data mapping,
- Cataloguing,
- Harvesting
Data librarian – skills

- enabling access,
- user experience,
- web interfaces,
- Graphical User Interfaces (GUI’s),
- Digital Object Identifiers (DOI’s),
- back up,
- version control,
- naming conventions,
Data librarian – skills

IT skills

- XML,
- database design,
- APIs,
- user centered design,
- natural language processing tools
- the Internet of things
The Bridge of Knowledge

Business
Researchers
Information society

open science

Most Wiedzy

publications
manuscripts
grants
projects
R&D teams
laboratories
research equipment
inventions
patents
trademarks
offers for business
The Bridge of Data

Multidisciplinary Open System Transfering Knowledge stage II Open Research Data

The goal was to create **Competence Center** which provides expertise and support including trainings among scholars about different aspects of Open Science and face to face consultations

and to design and build a **platform dedicated to research data** generated at the **three most important universities of Pomerania**, which will increase accessibility, coherence and reuse of science, knowledge and technology resources

co-financed by the European Regional Development Fund
Open Data Repository – mostwiedzy.pl

• Collect and store datasets from Gdańsk Tech, GUMed and UG
• Indexing datasets in Google Data, Web of Science Data Citation Index and other services
• Technological innovations such as hosting the project on the private computing cloud and storing the data on the Ceph Object Storage.
• NoSQL database – ElasticSearch.
• Repository allows researchers to perform Big Data Analysis by the Apache Zeppelin GUI on the supercomputer Tryton (40,000 cores, 1.5 PFLOPS).
Open Science Competence Center

- Assistance and on-site tailoring training among researchers from all scientific disciplines that include Data Management Plan, open licensing or metadata standards
- Workshops regarding different aspects of Open Science as well as scholarly communication
- InfoKit regarding Open Research Data
- Metadata support
- Journals & Conference Proceedings indexing and publishing support
- Evaluation & Bibliometric support
Data steward
BRINGING CONTROL TO CHAOS

Data stewar

Data Quality

Security

Privacy

Data Management
Data services roles and responsibilities
Open Science Competence Center - trainings

- Overview of open research data,
- Data Management Plans (national and EU grant applications),
- Legal support (data licensing, data protection, reusing data),
- Using the Bridge of Data Repository (depositing datasets),
- FAIR meta(data),
- Plan S (implications and requirements).
Project "MOST DANYCH"
is co-financed by the European Regional Development Fund
within the Operational Programme Digital Poland 2014-2020
Thank you