Data and metadata EDCH

Bibliothèque de l’EPFL

Why should you care?

Considering primary data as a research output with its own value is a general trend (started long ago in other fields, such as astrophysics):

- It supports research transparency and reproducibility
- Data science relies on the availability of data

Also :

- Creating and organizing data is always a part of the research process
- Well-organized data makes your research easier to follow (including for yourself in 12 months or more…)
- Journals, funding agencies and universities increasingly require to make the primary data that supports scientific publications available (to reviewers, to other researchers, to the greater public)

What is research data ?

«Data that are used as primary sources to support technical or scientific enquiry, research, scholarship, or artistic activity, and that are used as evidence in the research process and/or are commonly accepted in the research community as necessary to validate research findings and results. All other digital and non-digital content have the potential of becoming research data. Research data may be experimental data, observational data, operational data, third party data, public sector data, monitoring data, processed data, or repurposed data.»

https://codata.org/rdm-terminology/research-data/

Older examples from days 1 and 2

- Data journals : as seen on day 1, but the concept can be traced further back (*Journal of Physical and Chemical Reference Data* launched in 1972, for example)
- Specialized databases with citable records, i.e. CSD

Alain Borel

Data repositories : examples

- https://www.zenodo.org/communities/epfl (EPFL collection in a free repository operated by CERN)
- https://yareta.unige.ch/home/search?search=search%3Dchemistry (University of Geneva)

What research data should you consider?

- Minimum: experimental or computational data used to produce a published result
- Ideally, all software used to process the data during the analysis
- Intermediate/processed data if the workflow is difficult/expensive to reproduce

… and more if you want/can

What is metadata ?

«Literally, “data about data”; data that defines and describes the characteristics of other data, used to improve both business and technical understanding of data and data-related processes. Business metadata includes the names and business definitions of subject areas, entities and attributes, attribute data types and other attribute properties, range descriptions, valid domain values and their definitions. Technical metadata includes physical database table and column names, column properties, and the properties of other database objects, including how data is stored. Process metadata is data that defines and describes the characteristics of other system elements (processes, business rules, programs, jobs, tools, etc.). Data stewardship metadata is data about data stewards, stewardship processes and responsibility assignments.»

https://codata.org/rdm-terminology/metadata/

The value of raw data without any kind of description is close to zero.

No, really, what is metadata?

Information about the data : creators, methods, creation time, purpose… recorded in :

- File and folder names
- README.txt and similar files
- comments, tags or other annotations inside data files (the NMReDATA uses such an approach https://dx.doi.org/10.1002/mrc.4737)
- more specialized formats (for example https://zenodo.org/record/4572642/export/dcite4)

Problem : few modern and mature standards in chemistry, work in progress https://iupac.org/what-we-do/digital-standards/

Some interesting developments https://www.allotrope.org/allotrope-framework

The FAIR principles

- ***F* indable**
  Data and metadata are easy to find by both humans & computers.
  - Use metadata
  - Deposit (meta)data in repository/registry
  - Assign a persistent identifier (eg. DOI, HANDL, URN)

- ***A* ccessible**
  Machines & humans can readily access or download (meta)data.
  - As-open-as-possible access to your data (licensing, …)
  - Services with user-friendly interfaces
  - Leave the metadata available after data deletion

- ***I* nteroperable**
  Data from different datasets are ready to be exchanged or combined.
  - Use open file format(s), whenever possible
  - Use standardized vocabularies/tags
  - Use cross-linking as much as possible

- ***R* eusable**
  (Meta)data are easily replicated / combined in future research.
  - Attach standardized license to your data (CC, GPL, …)
  - Capture provenance information as precisely as possible

What does it look like?

https://doi.org/10.26037/yareta:526n2fl7gbe4jibfdcpbmtnpem

- ./data
- ./data/f03e6c3e-aae2-4130-83af-552b785123f9
- ./data/f03e6c3e-aae2-4130-83af-552b785123f9/researchdata
- ./data/f03e6c3e-aae2-4130-83af-552b785123f9/researchdata/IR
- ./data/f03e6c3e-aae2-4130-83af-552b785123f9/researchdata/Optical Properties
- ./data/f03e6c3e-aae2-4130-83af-552b785123f9/researchdata/Calculations
- ./data/f03e6c3e-aae2-4130-83af-552b785123f9/researchdata/Calculations/[(NHC)AuCl]
- ./data/f03e6c3e-aae2-4130-83af-552b785123f9/researchdata/Calculations/[(NHC)AuCl]/.olex
- ./data/f03e6c3e-aae2-4130-83af-552b785123f9/researchdata/Calculations/[(NHC)AuCl]/.olex/originals
- ./data/f03e6c3e-aae2-4130-83af-552b785123f9/researchdata/Calculations/[(NHC)Rh(cod)Cl]
- ./data/f03e6c3e-aae2-4130-83af-552b785123f9/researchdata/Calculations/[(NHC)Rh(cod)Cl]/.olex
- ./data/f03e6c3e-aae2-4130-83af-552b785123f9/researchdata/Calculations/[(NHC)Rh(cod)Cl]/.olex/originals
- ./data/f03e6c3e-aae2-4130-83af-552b785123f9/researchdata/Calculations/(NHC)Se
- ./data/f03e6c3e-aae2-4130-83af-552b785123f9/researchdata/Calculations/[(NHC)Au(C`^N^C)]
- ./data/f03e6c3e-aae2-4130-83af-552b785123f9/researchdata/Calculations/[(NHC)Au(C`^N^C)]/.olex
- ./data/f03e6c3e-aae2-4130-83af-552b785123f9/researchdata/Calculations/[(NHC)Au(C`^N^C)]/.olex/originals
- ./data/f03e6c3e-aae2-4130-83af-552b785123f9/researchdata/NMR
- ./data/f03e6c3e-aae2-4130-83af-552b785123f9/researchdata/NMR/2H.PF6 (600MHz)
- ./data/f03e6c3e-aae2-4130-83af-552b785123f9/researchdata/NMR/2H.PF6 (600MHz)/HSQC 1JC-H
- ./data/f03e6c3e-aae2-4130-83af-552b785123f9/researchdata/NMR/2H.PF6 (600MHz)/HSQC 1JC-H/pdata

What does it look like? (II)

```
./dlcm.xml
./data/f03e6c3e-aae2-4130-83af-552b785123f9/dlcm.xml
./data/f03e6c3e-aae2-4130-83af-552b785123f9/researchdata/IR/[(NHC)Rh(CO)2Cl]_reactIR.icIR
./data/f03e6c3e-aae2-4130-83af-552b785123f9/researchdata/Optical Properties/Optical Properties.xlsx
./data/f03e6c3e-aae2-4130-83af-552b785123f9/researchdata/Calculations/[(NHC)AuCl]/.olex/
NHC_Cyclohexyl_P-DMQA-Me-Pr_Au1-complex_aR_optimized.phil
./data/f03e6c3e-aae2-4130-83af-552b785123f9/researchdata/Calculations/[(NHC)AuCl]/.olex/
NHC_Cyclohexyl_P-DMQA-Me-Pr_Au1-complex_aS_optimized.phil
./data/f03e6c3e-aae2-4130-83af-552b785123f9/researchdata/Calculations/[(NHC)AuCl]/.olex/originals/
NHC_Cyclohexyl_P-DMQA-Me-Pr_Au1-complex_aS_optimized.xyz
./data/f03e6c3e-aae2-4130-83af-552b785123f9/researchdata/Calculations/[(NHC)AuCl]/.olex/originals/
NHC_Cyclohexyl_P-DMQA-Me-Pr_Au1-complex_aR_optimized.xyz
./data/f03e6c3e-aae2-4130-83af-552b785123f9/researchdata/Calculations/[(NHC)AuCl]/.olex/
NHC_Cyclohexyl_P-DMQA-Me-Pr_Au1-complex_aR_optimized.hist
./data/f03e6c3e-aae2-4130-83af-552b785123f9/researchdata/Calculations/[(NHC)AuCl]/.olex/
NHC_Cyclohexyl_P-DMQA-Me-Pr_Au1-complex_aS_optimized.hist
./data/f03e6c3e-aae2-4130-83af-552b785123f9/researchdata/Calculations/[(NHC)AuCl]/.olex/
NHC_Cyclohexyl_P-DMQA-Me-Pr_Au1-complex_aS_optimized.metacif
./data/f03e6c3e-aae2-4130-83af-552b785123f9/researchdata/Calculations/[(NHC)AuCl]/.olex/
NHC_Cyclohexyl_P-DMQA-Me-Pr_Au1-complex_aR_optimized.metacif
./data/f03e6c3e-aae2-4130-83af-552b785123f9/researchdata/Calculations/[(NHC)AuCl]/.olex/
NHC_Cyclohexyl_P-DMQA-Me-Pr_Au1-complex_aS_optimized.odb
./data/f03e6c3e-aae2-4130-83af-552b785123f9/researchdata/Calculations/[(NHC)AuCl]/NHC_Cyclohexyl_P-
DMQA-Me-Pr_Au1-complex_aS_optimized.xyz
./data/f03e6c3e-aae2-4130-83af-552b785123f9/researchdata/Calculations/[(NHC)AuCl]/NHC_Cyclohexyl_P-
DMQA-Me-Pr_Au1-complex_aR_optimized.bmp
```

FAIR data vs. supplementary information

	S.I. human ?	S.I. machine ?
***F* indable** Data and metadata are easy to find by both humans & computers.	☑	☒
***A* ccessible** Machines & humans can readily access or download (meta)data.	☑	±
***I* nteroperable** Data from different datasets are ready to be exchanged or combined.	±	☒
***R* eusable** (Meta)data are easily replicated / combined in future research.	±	☒

Rights and duties of data creators

- **Federal Act on Data Protection**[1] **+ EU GDPR** : protection of the personal data of humans
- **Federal Act on Copyright and Related Rights**[2] : copyright on intellectual creations, incl. software
- **Directive 96/9/EC pf the European Parliament**[3] : legal protection of databases
- **Federal Act on the Federal Institutes of Technology, art. 37**[4] : rights on data and code in the federal institutes
  - With the exception of copyright, all other rights to intellectual property [..] shall belong to the two federal institutes of technology and the four research institutes within the ETH Domain
  - The exclusive right to use computer programs created [...] rest solely with the two federal institutes [...]
  - The two federal institutes of technology and the four research institutes within the ETH Domain may enter into binding arrangements with the holders of other categories of copyright for the assignment of those rights.

Rights and duties (II)

In summary : EPFL generally owns everything you produce, except copyright.
The point where copyright becomes applicable in not so clear for data.

Side note: a Data Management Plan (DMP) is more and more often required during research project proposals.

Réf.s :
[1] Confédération Suisse, 2023. *Loi fédérale sur la protection des données* (LPD)
[2] Confédération Suisse, 2020. *Loi fédérale sur le droit d'auteur et les droits voisins* (LDA)
[3] Union Européenne, 2020. *Directive 96/9/CE du Parlement Européen et du Conseil* (Directive 96/9/CE)
[4] Confédération Suisse, 2017. *Loi fédérale sur les écoles polytechniques fédérales* (Loi sur les EPF)

What licence for research data?

1. If you are re-using existing data, check if its license impacts your choice
2. If not, you are free to decide, but some licensing terms can make more sense than others depending on the context. In particular, licensing has different implications for code and other data types. The Creative Commons are a good, systematic scheme for standard document licences, even if their validity for more general data is an open question.

data

- CC-BY for document-like data (reports, images, sound, videos…) : most Open Access-like
- CC-0 for databases or unstructured data : best for machine learning & similar use cases
- CC-BY-SA maybe ; avoid -NC and -ND

code

- CC-BY also recommended by OpenAIRE (i.e. European projects)
- The Creative Commons organization advise against using CC for software !
- Stick with the main free, open-source licences ?

sources : https://www.openaire.eu/research-data-how-to-license/
https://creativecommons.org/faq/#can-i-apply-a-creative-commons-license-to-software

How to find datasets, research software, etc.

Metadata can be copied, converted, etc. => building search engines is facilitated (in principle). A few solutions are available, but...

Mendeley Data:https://data.mendeley.com/research-data/?search=Hybrids%20of%20Cationic%20[4]Helicene%20and%20N-Heterocyclic%20Carbene ✗

Google Dataset Search:
https://datasetsearch.research.google.com/search?query=Hybrids%20of%20Cationic%20%5B4%5DHelicene%20and%20N-Heterocyclic%20Carbene&src=0 ✗

OpenAIRE Explorer: https://explore.openaire.eu/search/find?fv0=Dataset%20for%20Hybrids%20of%20Cationic%20%5B4%5DHelicene%20and%20N-Heterocyclic%20Carbene%20as%20Ligands%20for%20Complexes%20Exhibiting%20(Chir)Optical%20Properties%20in%20the%20Far%20Red%20Spectral%20Window&f0=q ✓

Clarivate Data Citation Index (not available at EPFL): https://clarivate.com/products/scientific-and-academic-research/research-discovery-and-workflow-solutions/webofscience-platform/data-citation-index/