b

UNIVERSITÄT
BERN

How to efficiently document research data

Olga Churakova and Gero Schreier

Open Science Team, University Library Bern, openscience@unibe.ch

https://creativecommons.org/licenses/by/4.0/deed.en

Open Science
Research Data Management Support

Head
Dr. Dirk Verdicchio

Business, Economics and Social Sciences
Jennifer Morger

Humanities, Law and Theology
Digital Humanities
Ursula Loosli

Humanities, Law and Theology
Dr. Gero Schreier

Medicine, Insel, Vetsuisse
Dr. habil. Olga Churakova

Natural Sciences
Dr. Martin Wegmann

Data Science IT
Dr.-Ing. Federico Grasso Toro

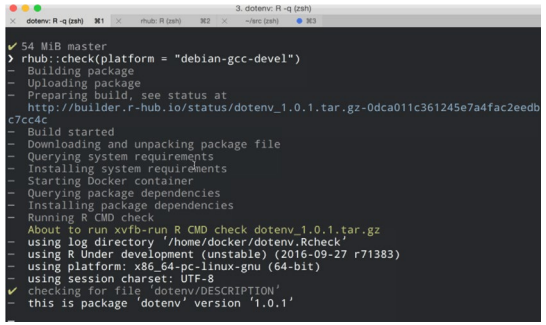
Human Sciences
Dr. Christine Krebs

Outline

Part 1:
- Documentation: Why, what, how?
- Documentation tools: what to consider?

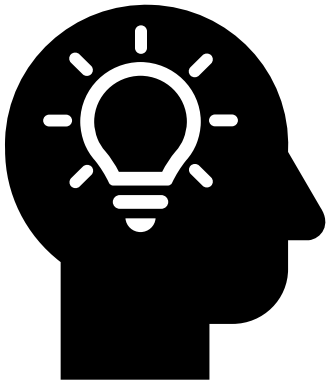
Part 2:
- What is metadata?
- What is controlled vocabularies?

Part 3:
- Support

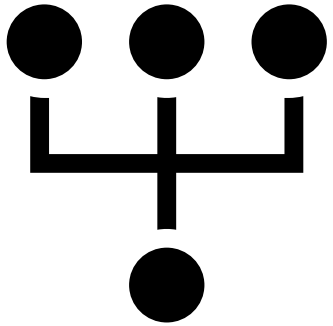


https://www.perkinelmer.com/de/

Part 1: Documentation



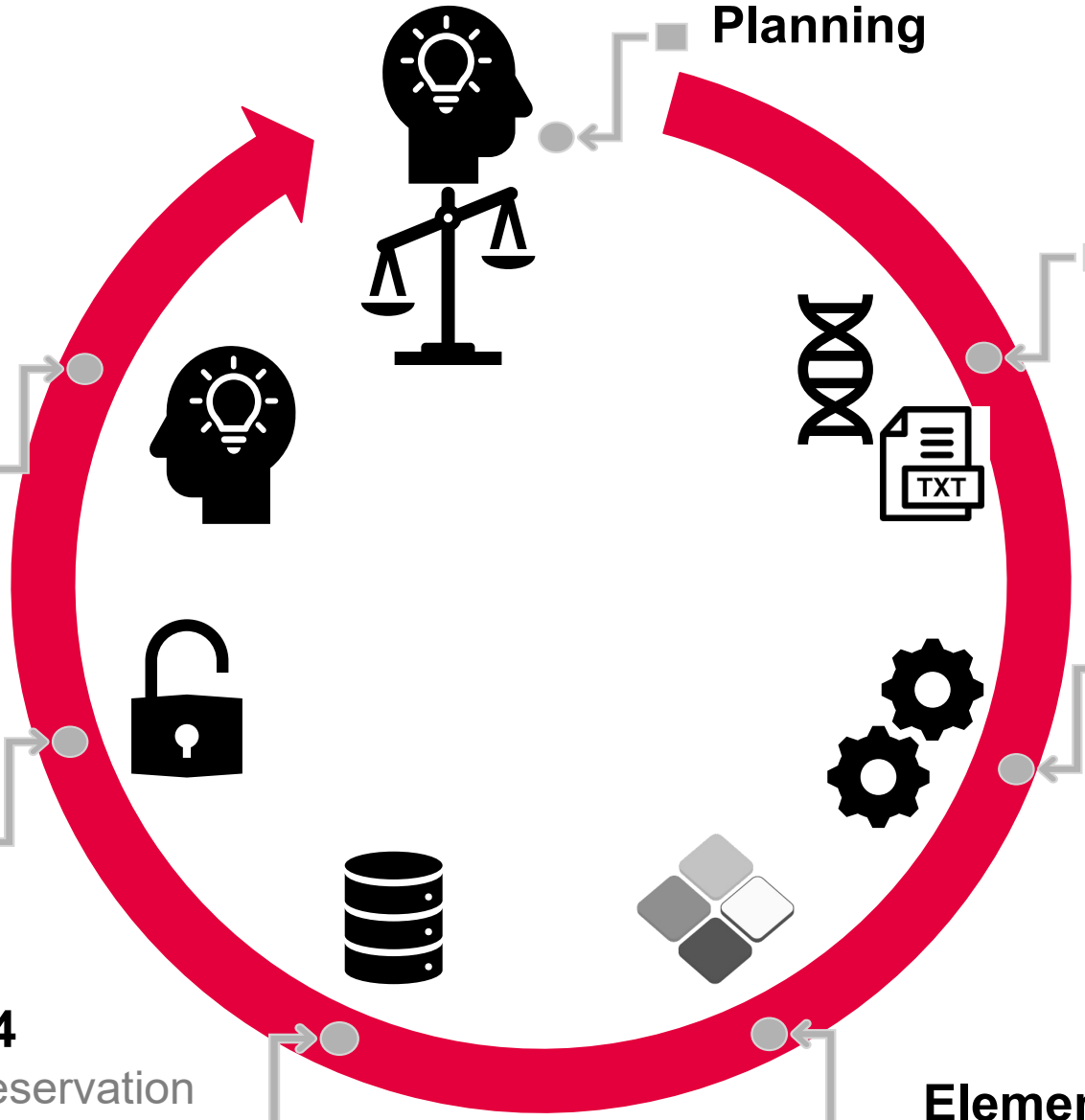
Why?



What?

How?

Why?



Planning

Element 1
- Data type
- Format
- Data description
- Data size
- File convention
- **Data Documentation**

Element 2
- Tools
- Software
- Code

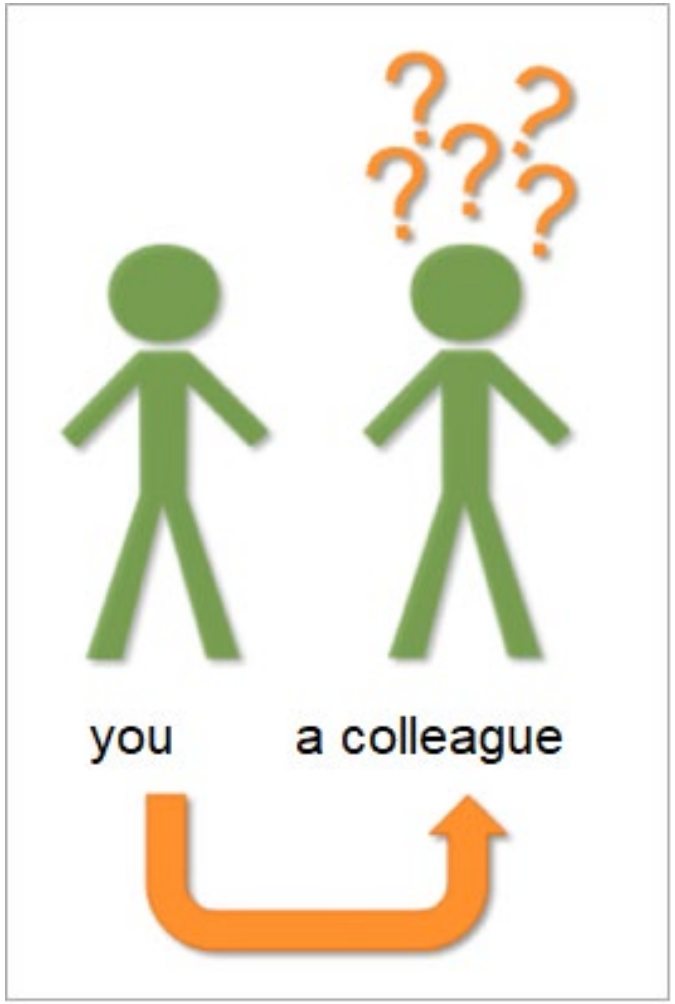
Element 3
- Metadata & Standards

Element 4
- Data preservation
- Access
- Timeline

Element 5
- Access
- Distribution
- Reuse

Element 6
- DMP
- Updates
- Monitor
- Publishing

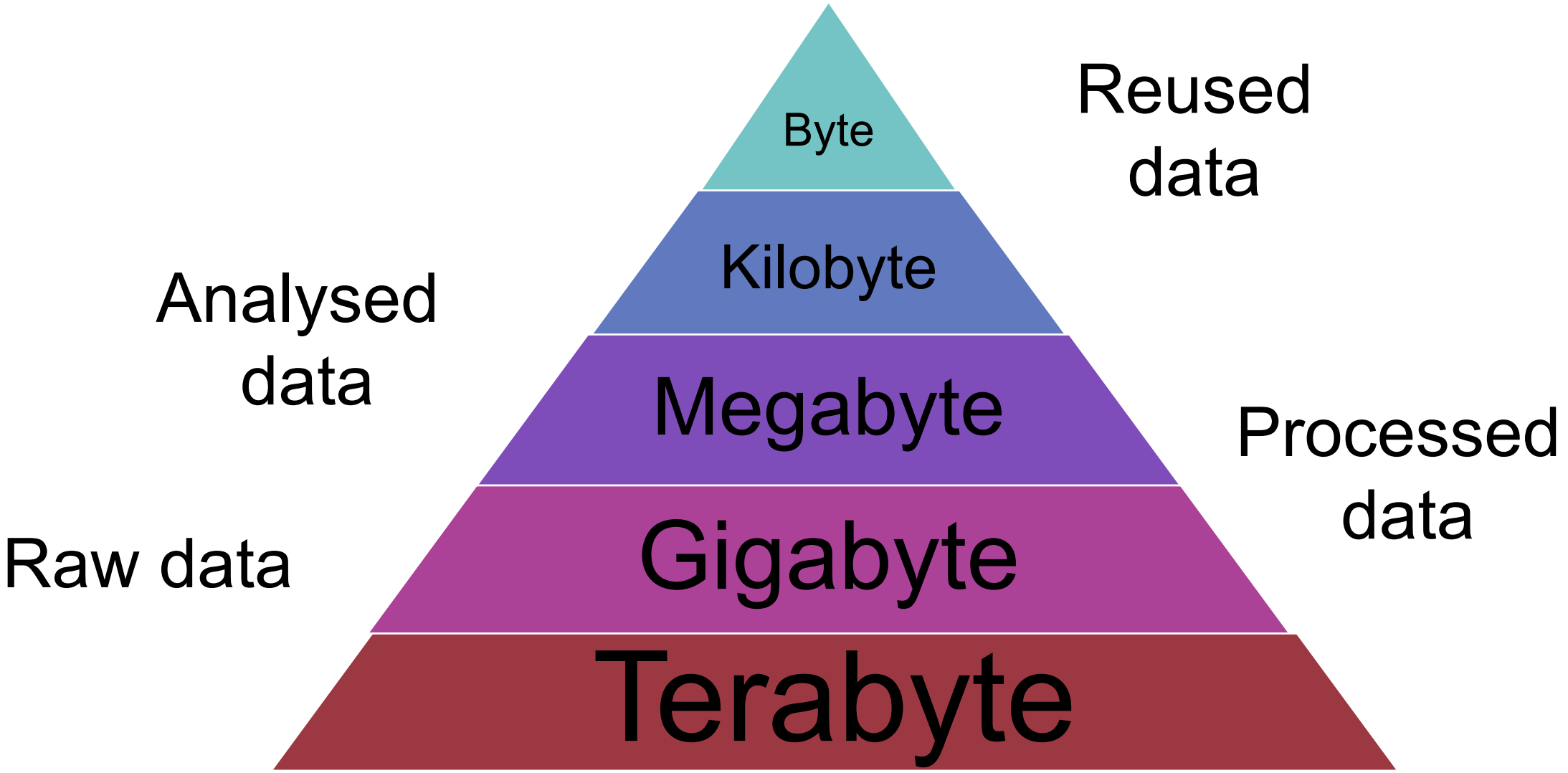
Why to document data?



Christine Malinowski, Data Management: File Organization, LIT Libraries, 24.07.2019, CC BY.

- **Who** created the data?
- **What** is the content of the data?
- **Why** were the data developed?
- **Where** is it geographically located?
- **When** were the data created?
- **How** were the data developed?

What kind of data should be documented?



What to consider?
Recommendations and Best Practices

- Data documentation based on FAIR principles
- Code of conduct for scientific integrity
- Open Science UB UniBE recommendation on research data documentation (PDF, 141KB)
- Storage of data documentation (10-15 years)
- Preparing data documentation in relation to research data

Swiss Academies of Art and Sciences

Go FAIR

What to consider?

- Start documentation during initiation project phase
- Create a separate documentation file, supplementary for the data
- Document consistently throughout the project
- Write your documentation in English (or common working language in your group)
- Documentation in electronic form makes searching and sharing information easier and saves time ⟶ documentation tools

What to consider?

File-level Documentation: Naming

- Creation date (YYYY-MM-DD)
- Project reference/name
- Description of the content
- Name of creator (initials or whole name)
- Name of research team/department
- Version number (v01)

What to consider?

Project-level documentation

Purpose of data creation?	What does the dataset contain?	How was data collected?	Who collected the data, when and in frame of which project?
Primary data source or raw data	How was the data processed?	How was data quality assured?	How can data be accessed?

What to consider?

File-level Documentation: ReadMe

- **Who** created the data?
- **What** is the content of the data?
- **Why** were the data developed?
- **Where** is it geographically located (coordinates)?
- **When** were the data created?
- **How** were the data developed?

- **Information on variables**: labels, codes, classifications, units, reference to question numbers missing values, aggregations

Help text is included in angle brackets, and should be deleted>

<Lines can be multiplied or deleted if needed>

This DATASETNAMEreadme.txt file was generated on [YYYYMMDD] by [Name]

GENERAL INFORMATION
---
1. Title of dataset:

2. Contributor information:
<describe person/s involved, e.g., principal investigator, data creator, data submitter etc. and if needed, add information about an alternate contact person for questions>
Name:
Role/Function:
Institution:
Address:
Email:

3. Date of data collection:
<single date, range, approximate date. Suggested format YYYY-MM-DD>

4. Geographic location of data collection:
<where was data collected? E.g., site information, site name, latitude, longitude, elevation>

5. Keywords describing the subject of your dataset:
<It is recommended to use controlled vocabularies in the form 'term (identifier)', for example 'Aachen (http://vocab.getty.edu/page/tgn/7004799)'. Otherwise, you can also use free keywords.>

6. Information about funding sources that supported the collection of the data:
Funding agency name:
Grant number:

SHARING/ACCESS INFORMATION
---
1. Licenses/restrictions placed on the data:
<name licenses and eventually explain restrictions on your data>

2. Links to publications that cite or use the data:

3. Links to other publicly accessible locations of the data:

4. Links/relationships to additional data sets:

5. Was data derived from another source? yes/no.
<If yes, list source(s):>

6. Recommended citation for this dataset:

Example: Readme_Template_EN.txt (3KB)

What to consider?

File-level Documentation: Codebook

Describe

- Contents
- Structure
- Variable name
- Layout of data collections

Use

Tabular (raws, columns) / statistical data

WFHN RedCap Codebook

Variable / Field Name	Field Type	Field Label	Choices, Calculations, OR Slider Labels	Field Note	Required Field?
rti_id_number	text	RTI ID Number		Participant ID	
rti_watch_order	dropdown	RTI Watch Order	1, 1 | 2, 2 | 3, 3 | 4, 4 | 5, 5	Order of a given watch for the RTI participant	
rti_industry	text	RTI Industry			
rti_site	text	RTI Site			
rti_wave	text	RTI Wave			
rti_assign_date	text	RTI Assign Date			
rti_return_date	text	RTI Return Date			

What to consider?

Documentation Tools

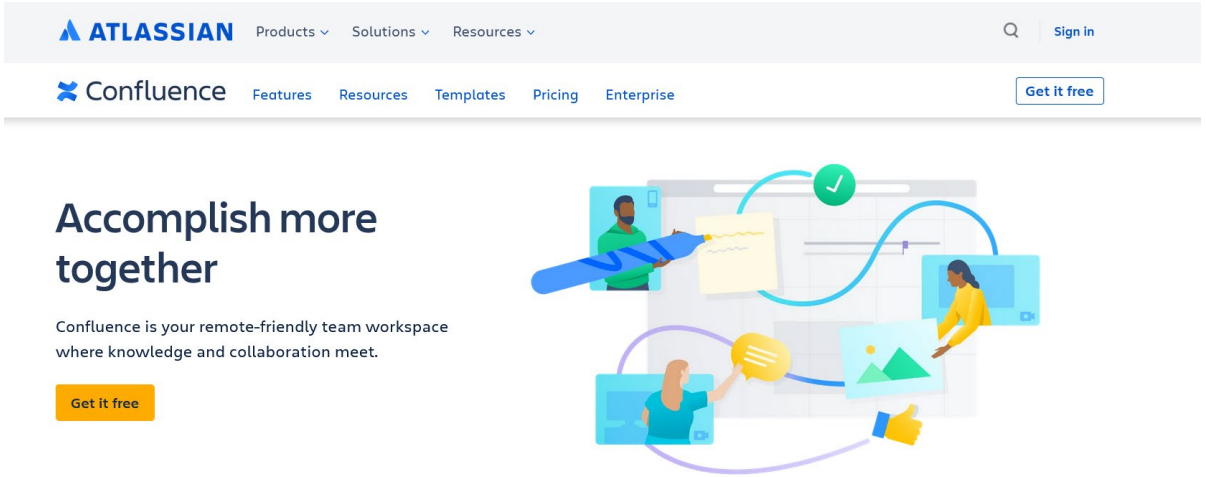
- What features do you need?
- Do you work on your own or should the tool work on different platforms?
- Free/open source or commercial product?
- Be careful when working with sensitive data!
- Export functions → will you be able to switch to another tool?
- Track changes (in forms, protocols)?
- Link to statistical software, computing plots and printing

https://www.perkinelmer.com/de/

How?
Documentation Tools

Project-level documentation:

- Confluence



https://www.atlassian.com/software/confluence

How?

Computational Notebooks

Jupyter Notebook

Open-source web application

- Create and share documents containing live code, equations, visualizations and narrative text
- Uses include:
  - Comments in scripts
  - Data cleaning and transformation
  - Numerical simulation and statistical modeling
  - Data visualization
  - Create tutorials and interactive manuals

Description from: Jupyter procect, https://jupyter.org/documentation;
J.M.Perkel: Why Jupyter is data scientists' computational notebook of choice DOI: 10.1038/d41586-018-07196-1

How?

Searchable
Protocol, sample, data set

Automated
Report generation, manuscript writing, or inventory tracking

Consistent
Results are accurate and reproduceable

Collaborative
Knowledge exchange among researchers and stakeholders, regulatory bodies, and funders

Traceable
Track experiments, samples, protocols, and results

https://www.dlcm.ch/resources/dlcm-eln

Part 1: Documentation

Summary

- Document your data earlier enough
- Documentation (e.g., study protocols, data collection instruments)
- Project-level documentation (e.g., Confluence)
- Data-level documentation
  - Naming conventions
  - ReadMe files
  - Codebooks (definition of variables., e.g., REDCap; secuTrial)
  - Data labels
- Electronic Laboratory Notebooks (ELN) and Laboratory Information Management System (LIMS, e.g., OpenBIS)

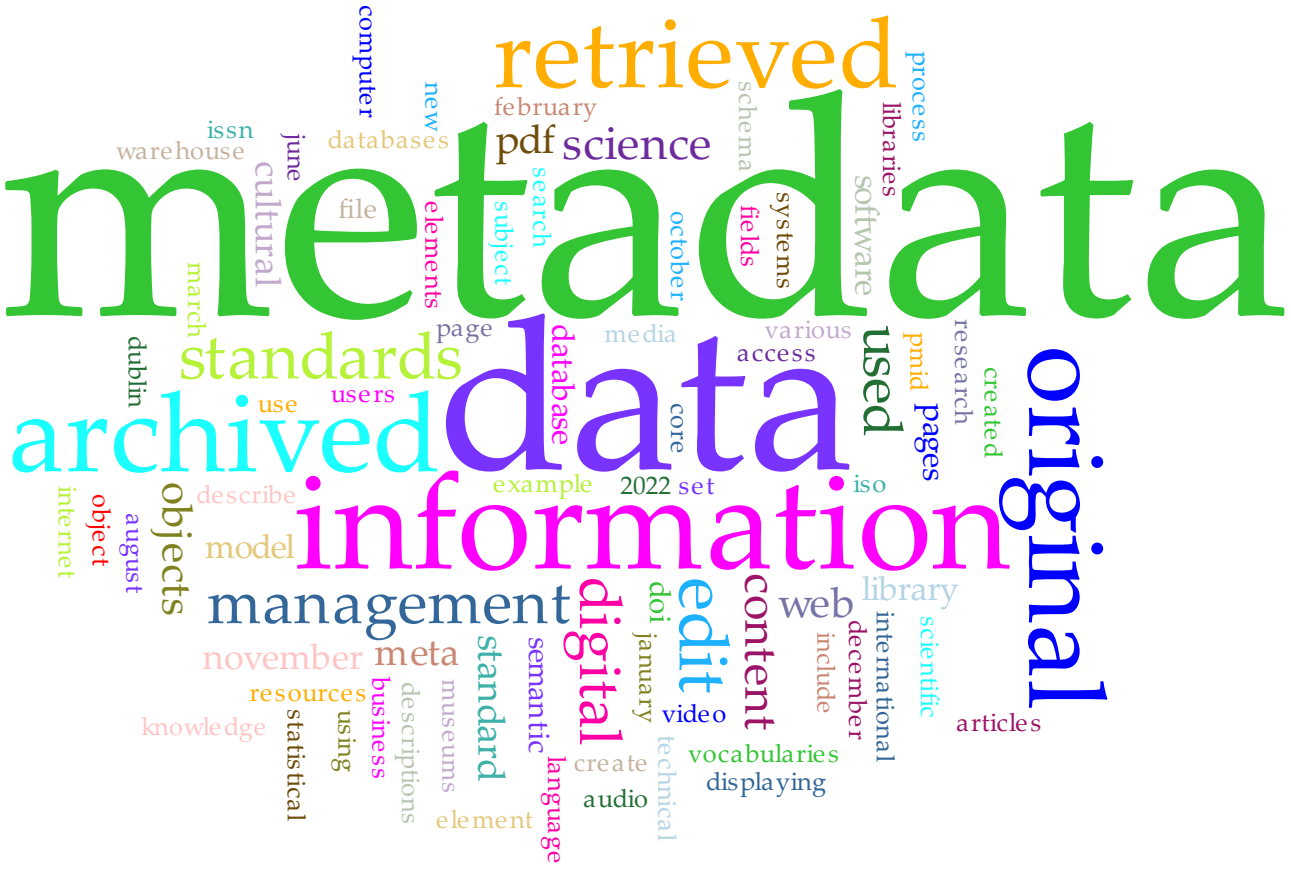
☑ Recommendation on research data documentation from the Open Science Team (PDF, 141KB)

Part 1: Documentation

Why, What and How?

Questions so far?

Part 2: Metadata

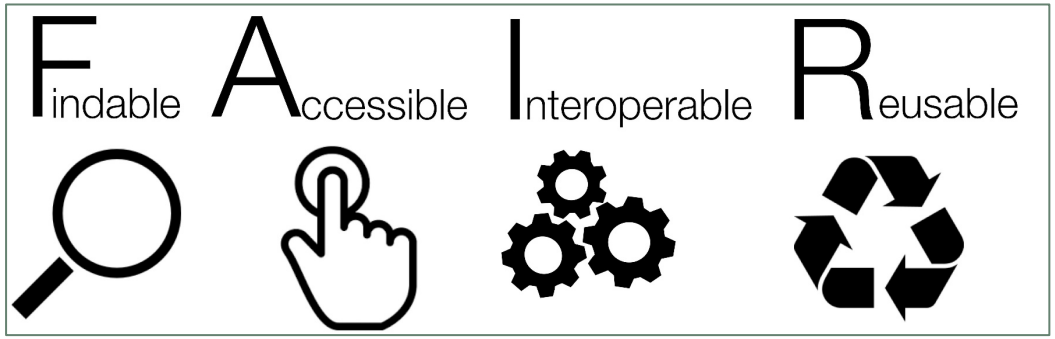


Wordcloud of https://en.wikipedia.org/wiki/Metadata, made with https://voyant-tools.org, Nov 14, 2023.

Metadata
Definitions

Metadata: a standardized, highly structured description of objects (including data).

Metadata schema / **standard**: the structure that determines which information is contained in a metadata record; corresponds to the fields of a repository input mask or labels of a record.



Metadata help to

- Make information machine-actionable
- Retrieve research data
- Structure large amounts of data
- Perform meta-analyses
- Clarify rights and re-use conditions
- Support archiving and preservation

Metadata
Example

Previous research highlighted how algorithms on social media platforms can be abused to disseminate disinformation. However, less work has been devoted to understanding the interplay between Facebook news curation mechanisms and propaganda content. To address this gap, we analyze the activities of RT (formerly, Russia Today) on Facebook during the 2020 U.S. presidential election. We use agent-based algorithmic auditing and frame analysis to examine what content RT published on Facebook and how it was algorithmically curated in Facebook News Feeds and Sea Results. We find that RT's strategic framing included the promotion of anti-Biden leaning content, with an emphasis on antiestablishment narratives. However, due to algorithmic factors on Facebook, individual agents were exposed to eclectic RT content without an overarching narrative. Our findings contribute to the debate on computational propaganda by highlighting the ambiguous relationship between government-sponsored media and Facebook algorithmic curation, which may decrease the exposure of users to propaganda and at the same time increase confusion.

```
59 <meta name="DC.identifier" content="https://boris.unibe.ch/188806/" />
60 <meta name="DC.title" content="Blame It on the Algorithm? Russian Government-Sp
61 <meta name="DC.creator" content="Kuznetsova, Elizaveta" />
62 <meta name="DC.creator" content="Makhortykh, Mykola" />
63 <meta name="DC.subject" content="000 Informatik, Wissen, Systeme" />
64 <meta name="DC.subject" content="070 Publizistische Medien, Journalismus, Verla
65 <meta name="DC.subject" content="300 Sozialwissenschaften, Soziologie, Anthropo
66 <meta name="DC.subject" content="320 Politikwissenschaft" />
67 <meta name="DC.description" content="Previous research highlighted how algorith
68 <meta name="DC.publisher" content="USC Annenberg" />
   <meta name="DC.date" content="2023" />
   <meta name="DC.type" content="info:eu-repo/semantics/article" />
71 <meta name="DC.type" content="info:eu-repo/semantics/publishedVersion" />
72 <meta name="DC.type" content="PeerReviewed" />
73 <meta name="DC.format" content="application/pdf" />
74 <meta name="DC.language" content="eng" />
75 <meta name="DC.rights" content="info:eu-repo/semantics/openAccess" />
76 <meta name="DC.relation" content="https://boris.unibe.ch/188806/1/18687-71175-1
77 <meta name="DC.source" content="Kuznetsova, Elizaveta; Makhortykh, Mykola (2023
78 <meta name="DC.bibliographicCitation" content="International journal of communi
79 <meta name="DC.relation" content="Kuznetsova, Elizaveta; Makhortykh, Mykola (20
80 <meta name="DC.relation" content="https://ijoc.org/index.php/ijoc/article/view/
81 <meta name="DC.relation" content="1932-8036" type="issn" />
```

Example: https://boris.unibe.ch/188806/

Metadata

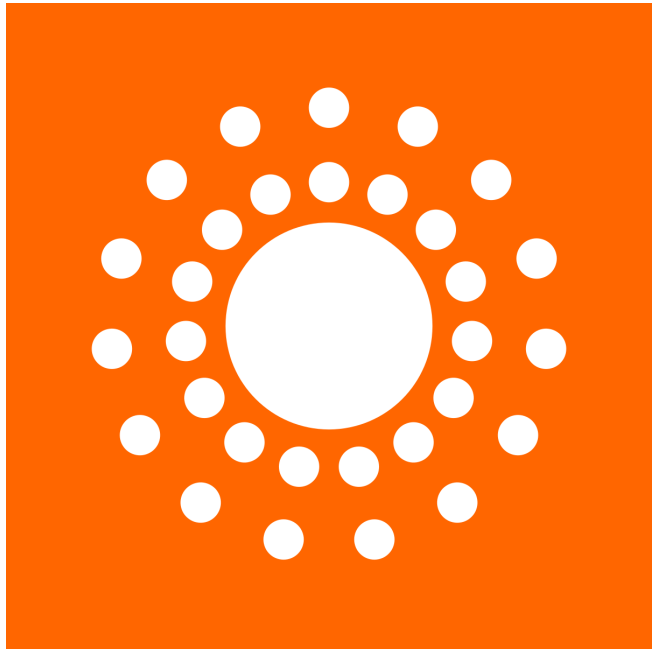
Metadata Type	Example Properties	Primary Uses
Descriptive metadata	Title Author Subject Genre Publication date	Discovery Display Interoperability
Technical metadata	File type File size Creation date/time Compression scheme	Interoperability Digital object management Preservation
Preservation metadata	Checksum Preservation event	Interoperability Digital object management Preservation
Rights metadata	Copyright status License terms Right holder	Interoperability Digital object management
Structural metadata	Sequence Place in hierarchy	Navigation
Markup languages	Paragraph Heading List Name Date	Navigation Interoperability

Jenn Riley NISO, 2017
www.niso.org

Metadata

Dublin Core: Key Metadata Elements

- Contributor
- Title
- Creator
- Date
- Description
- Type
- Rights
- Relation
- Publisher
- Format
- Language
- Subject
- Identifier
- Coverage
- Source



Agnew, 2005

Metadata
BORIS Portal

Based on Dublin Core elements

Title:	Daily suicides and ambient temperature data in Switzerland
Contributor(s):	Bär, Séverine Vicedo Cabrera, Ana Maria Bundo, Marvin Müller, Thomas de Schrijver, Evan
Affiliations:	Institute of Social and Preventive Medicine Institute of Social and Preventive Medicine Institute of Social and Preventive Medicine Institute of Social and Preventive Medicine
Contact:	Vicedo Cabrera, Ana Maria
Data Availability:	Open
Keyword(s):	epidemiology;public health;climate
Subject(s):	600 - Technology > 610 - Medicine & health
Description:	The dataset consists of daily number of suicides per Canton in Switzerland between 1995 and 2016 by age, sex and method of suicide. It also includes daily mean temperature per Canton.
URI:	http://hdl.handle.net/20.500.12422/71
DOI:	https://doi.org/10.48620/38
Rights:	cc-nc
Type:	Dataset
Appears in Collections:	Research Data

https://doi.org/10.48620/38

Metadata

Data Cite for research data

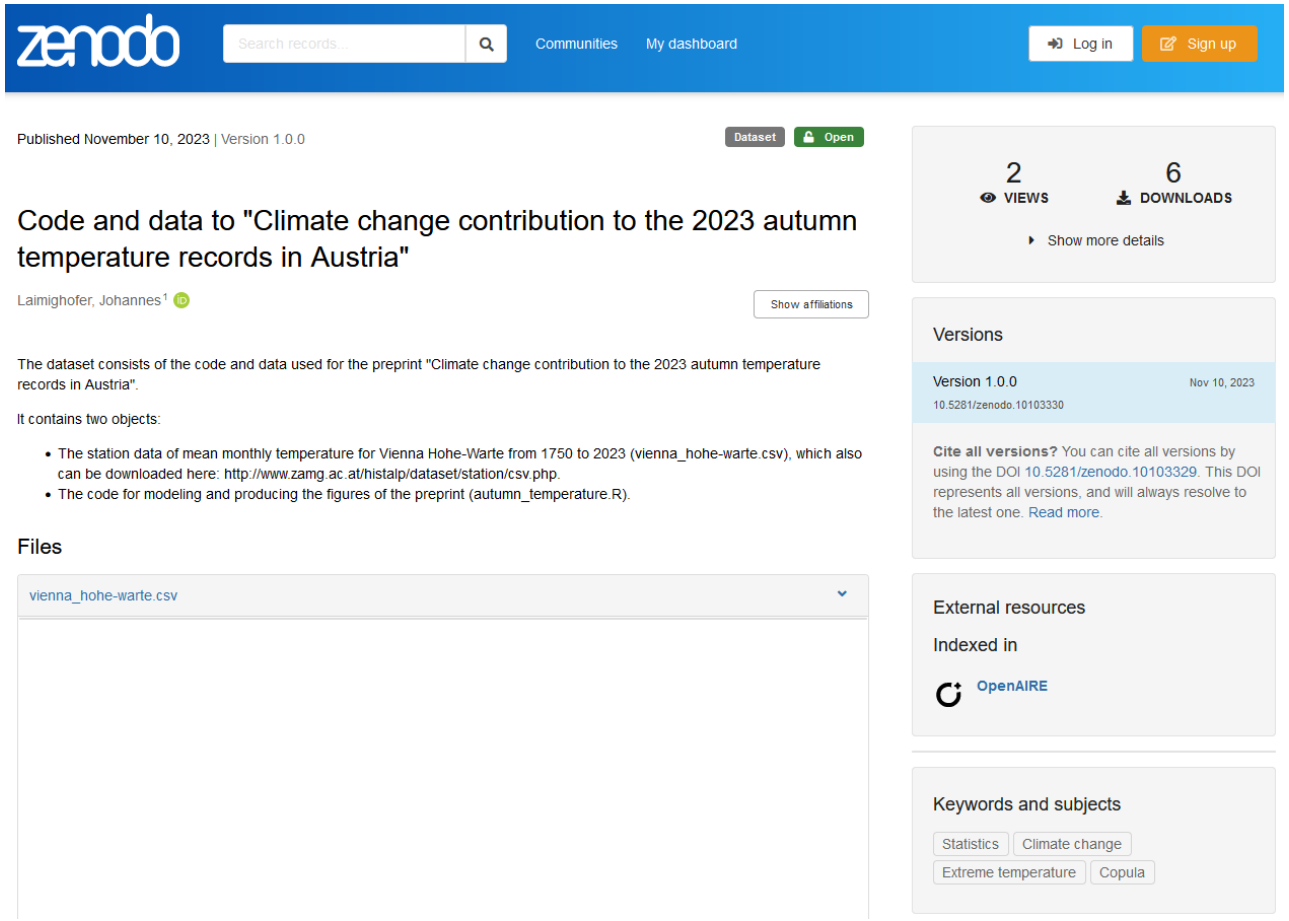
Mandatory	Recommended	Optional
Identifier	Subject	Language
Creator	Contributor	Alternate ID
Title	Date	Size
Publisher	Resource Type	Format
Publication year	Related identifier	Version
	Description	Rights
	GeoLocation	

DataCite
FIND, ACCESS, AND REUSE DATA

https://project-thor.readme.io/docs/datacite-metadata-schema

https://datacite.org/

Metadata
Data Cite for research data



Metadata
Controlled Vocabularies

"Controlled vocabularies provide a clearly defined terminology to catalog and retrieve information with control of synonyms and variant terms."
(RDM services, KU Leuven)

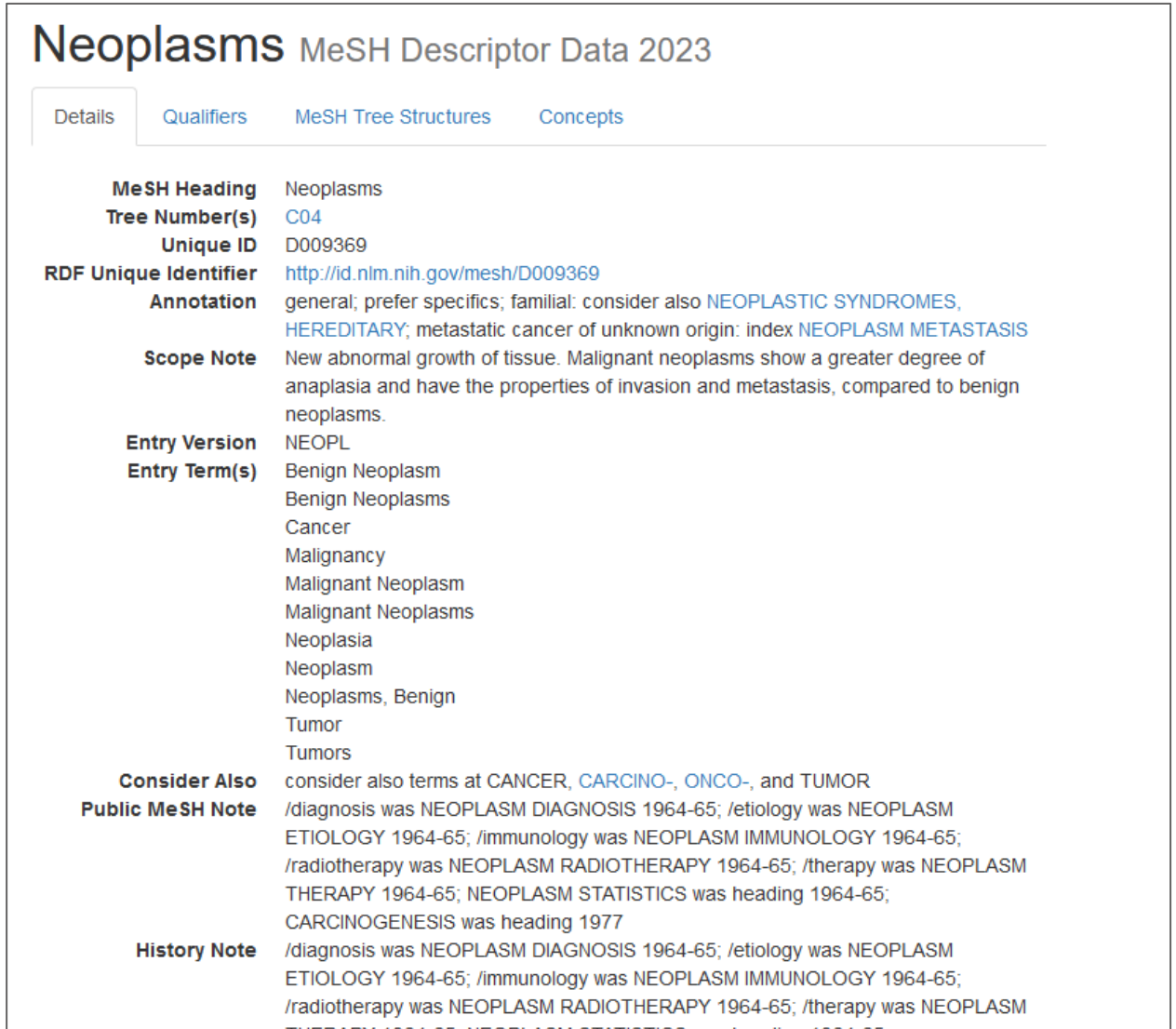
Controlled Vocabularies are a metadata content standard.

Common types include

- Subject headings lists (e.g. LCSH)
- Authority files
- Thesauri
- Ontologies

Metadata
Controlled Vocabularies

- They reduce ambiguity of terms (eg., "cancer")
- They provide a preferred form of a term in case of multiple variants (eg. "neoplasm")
- They help structure large amounts of data and make information quicker to retrieve (eg., library catalogs or databases)
- They provide unique identifiers that make terms machine-actionable

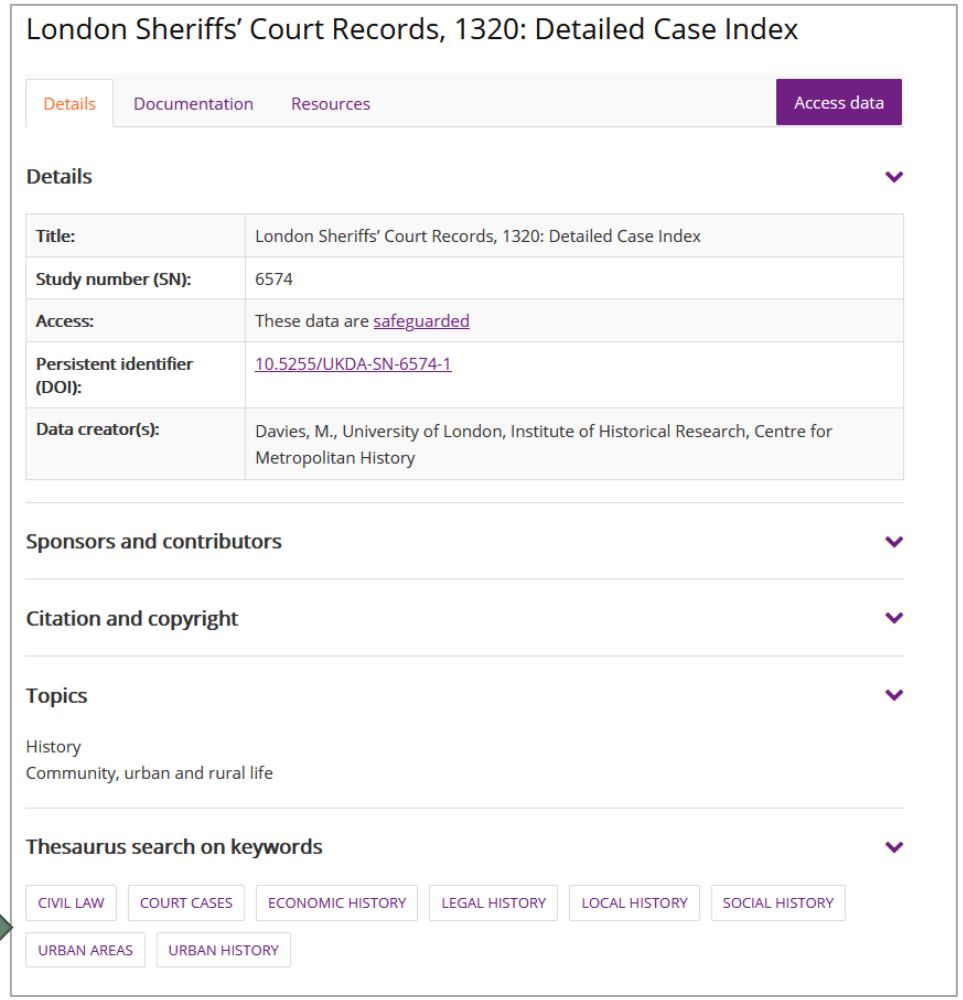


Neoplasms MeSH Descriptor Data 2023

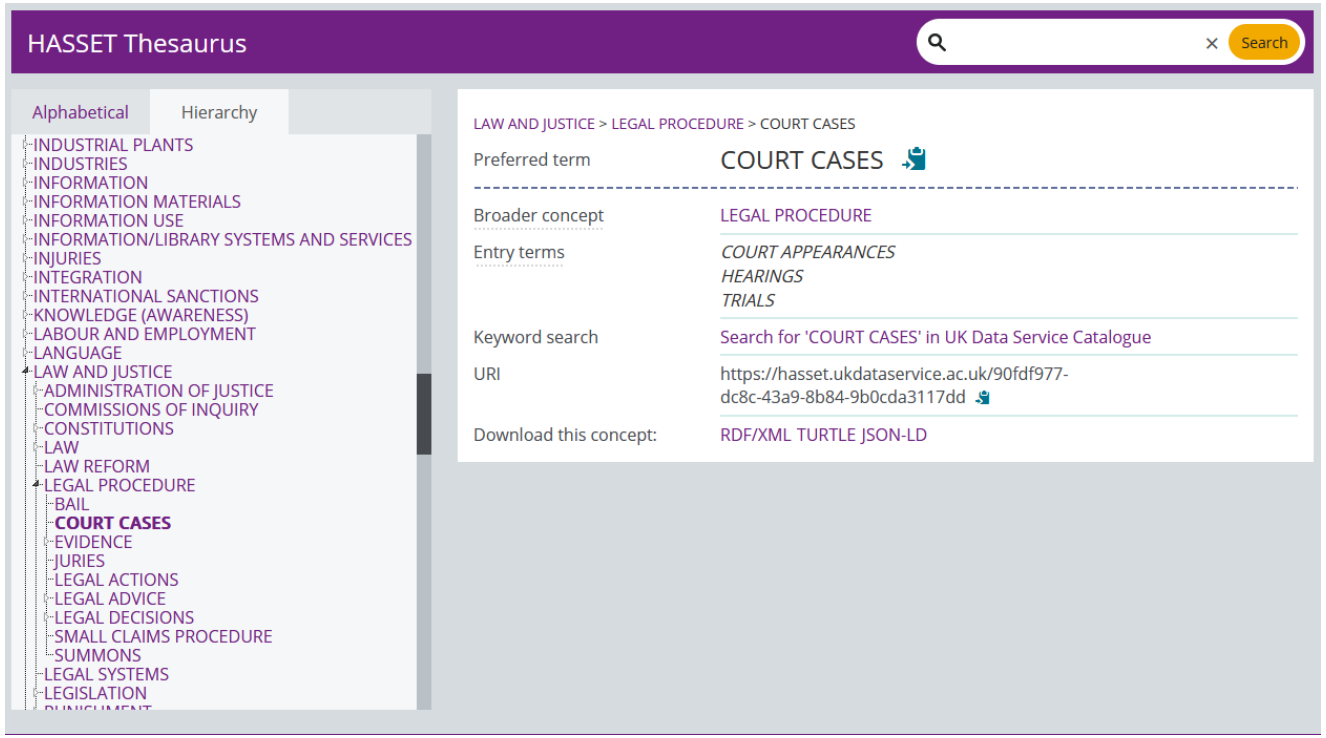
Details | Qualifiers | MeSH Tree Structures | Concepts

MeSH Heading	Neoplasms
Tree Number(s)	C04
Unique ID	D009369
RDF Unique Identifier	http://id.nlm.nih.gov/mesh/D009369
Annotation	general; prefer specifics; familial: consider also NEOPLASTIC SYNDROMES, HEREDITARY; metastatic cancer of unknown origin: index NEOPLASM METASTASIS
Scope Note	New abnormal growth of tissue. Malignant neoplasms show a greater degree of anaplasia and have the properties of invasion and metastasis, compared to benign neoplasms.
Entry Version	NEOPL
Entry Term(s)	Benign Neoplasm Benign Neoplasms Cancer Malignancy Malignant Neoplasm Malignant Neoplasms Neoplasia Neoplasm Neoplasms, Benign Tumor Tumors
Consider Also	consider also terms at CANCER, CARCINO-, ONCO-, and TUMOR
Public MeSH Note	/diagnosis was NEOPLASM DIAGNOSIS 1964-65; /etiology was NEOPLASM ETIOLOGY 1964-65; /immunology was NEOPLASM IMMUNOLOGY 1964-65; /radiotherapy was NEOPLASM RADIOTHERAPY 1964-65; /therapy was NEOPLASM THERAPY 1964-65; NEOPLASM STATISTICS was heading 1964-65; CARCINOGENESIS was heading 1977
History Note	/diagnosis was NEOPLASM DIAGNOSIS 1964-65; /etiology was NEOPLASM ETIOLOGY 1964-65; /immunology was NEOPLASM IMMUNOLOGY 1964-65; /radiotherapy was NEOPLASM RADIOTHERAPY 1964-65; /therapy was NEOPLASM

Metadata
Controlled Vocabularies



10.5255/UKDA-SN-6574-1



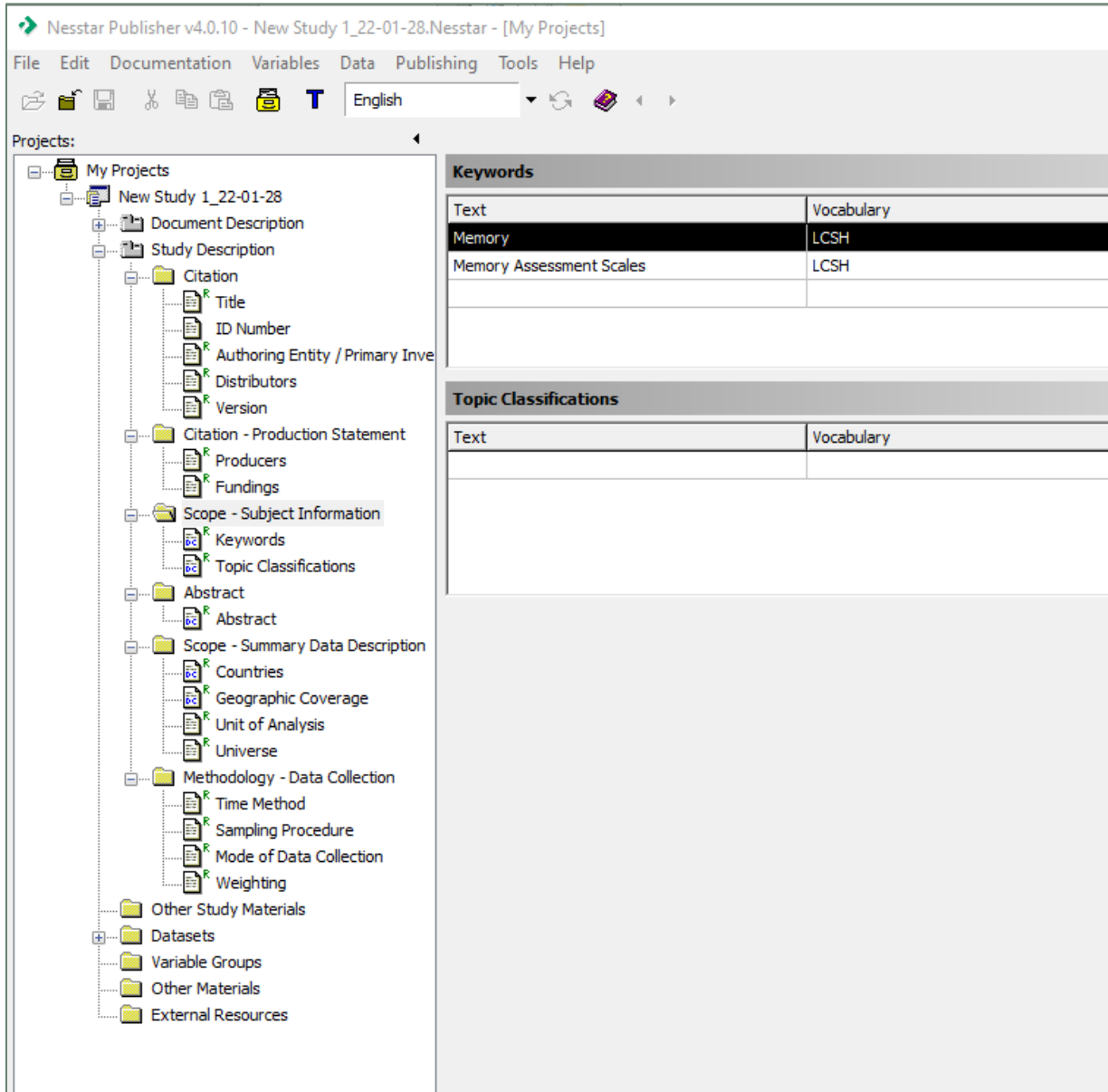
https://hasset.ukdataservice.ac.uk/hasset/en/page/90fdf977-dc8c-43a9-8b84-9b0cda3117dd

Metadata

How to use?

During research project

- Generate metadata that are specific to your field of study, data collection method, variables, etc.
- Metadata can be embedded in files or stored in separate file
- Some instruments / software generate metadata automatically
- Use a metadata tool
- Start as early as you can!

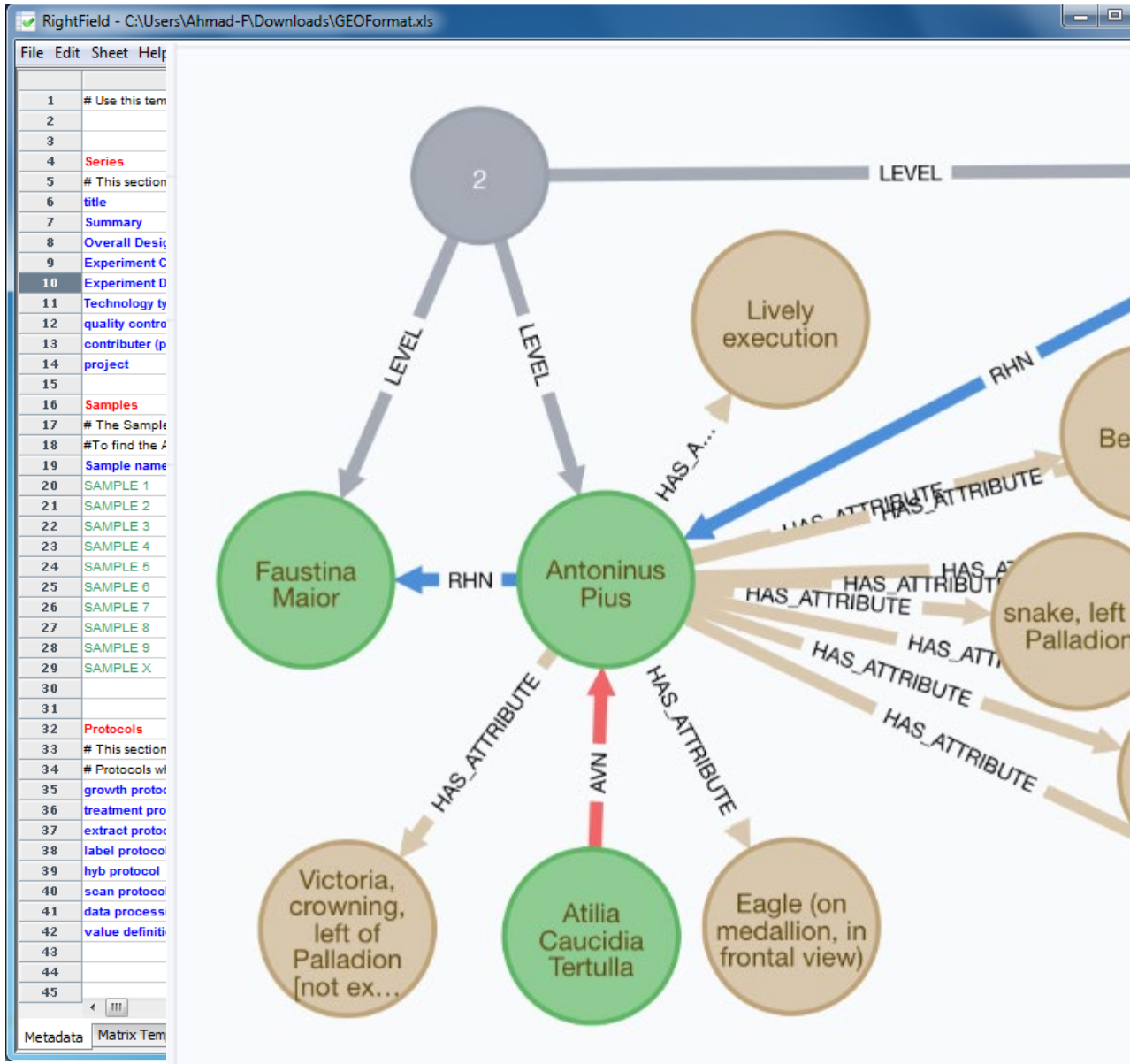


Metadata

How to use?

Controlled vocabularies

- Can be used in spreadsheets to make data structure more consistent
- Can provide unique identifiers that are needed to build databases



https://rightfiel

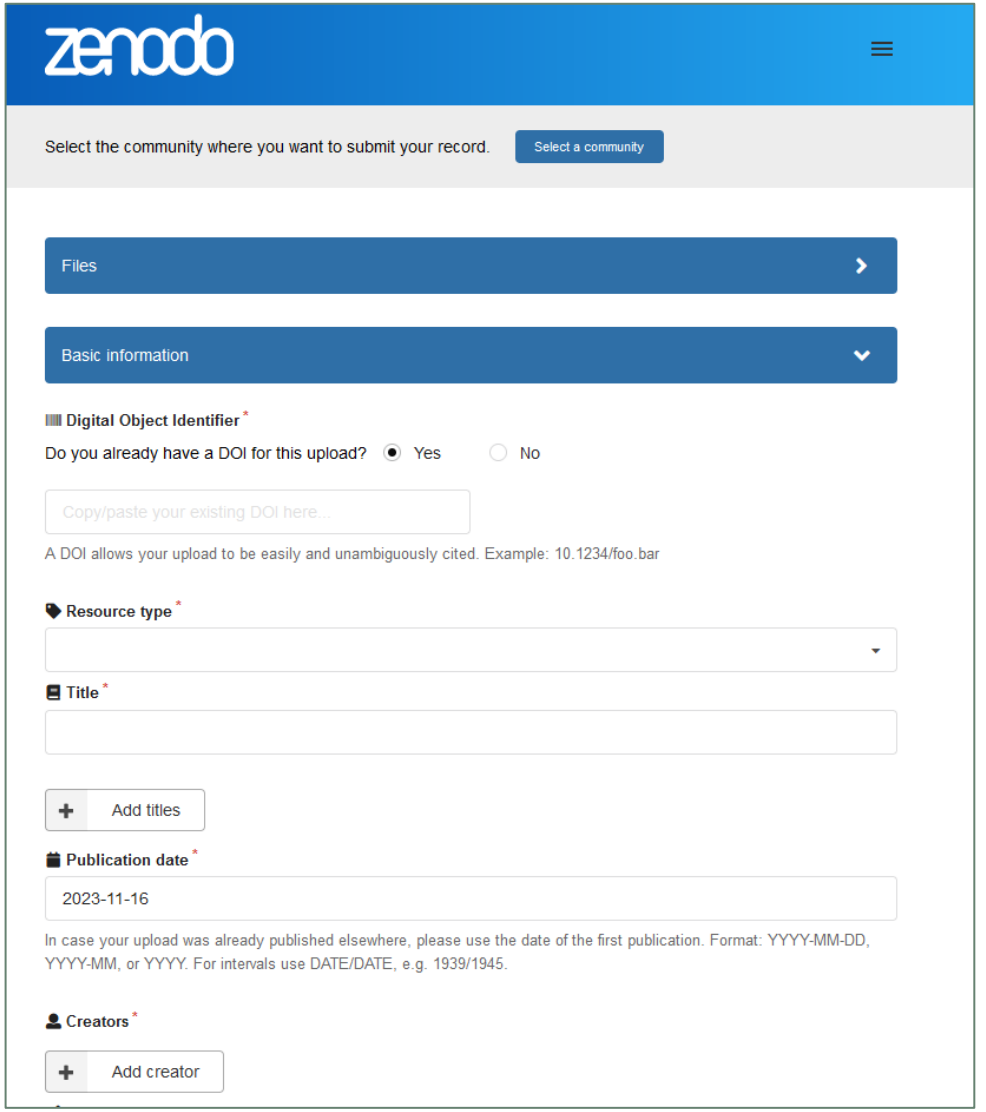
https://mittelalter.hypotheses.org/25591

Metadata

How to use?

When publishing data

- Generate metadata for citation and retrieval when uploading data to a repository (eg. Zenodo or BORIS Portal)
- Fill in repository input mask
- Include files with specific metadata (if applicable)
- Fill in as many fields as you can!



Metadata standards

Standards for many fields

- Darwin Core (Biodiversity data)
- Ecological Metadata Language (EML)
  Environmental data initiative (EDI)
- Climate and forecast (CF)
- Clinical Data Interchange Standards Consortium (CDISC)
- Data Documentation Initiative (DDI)
  Codebook or Lifecycle (Social Sciences)
- … and more!

Metadata standards catalogs:

- https://rdamsc.bath.ac.uk
- https://fairsharing.org (filter for "standards")

Tools for generating metadata:

- Dublin Core Generator
- Data Cite generator
- Cedar Workbench (Life Sciences / multidisciplinary)
- DCC List of metadata tools

Controlled vocabularies

Standards for many fields

Examples:

- Medical Subject Headings (MeSH)
- Astronomy Thesaurus
- Thesaurus for Economics
- Getty Thesaurus of Geographic Names: The Getty Research Institute
- DDI Vocabularies: Social Sciences Research)

Directories: Find a CV in your field!

- **L**inked **o**pen **ter**minology **re**sources
- Fairsharing.org
- Basic register of thesauri, ontologies & classifications
- JISC: Directory of Metadata Vocabularies
- Linked Open Vocabularies

Metadata standards

Summary

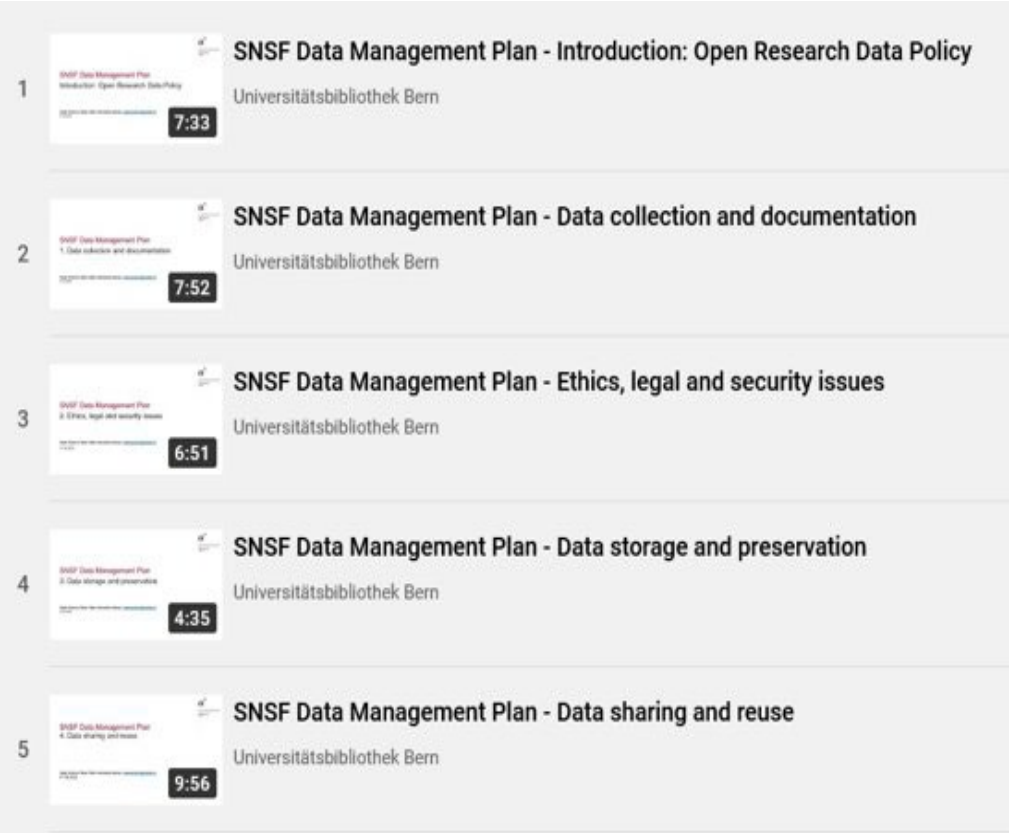
1) Use standardized metadata to
   - increase findability of your data
   - make data citable => get credit for your data
   - make data more FAIR

2) Assign standardized metadata to your dataset
   - automatically (instruments, software)
   - using a metadata tool
   - filling out all the fields in the repository input mask

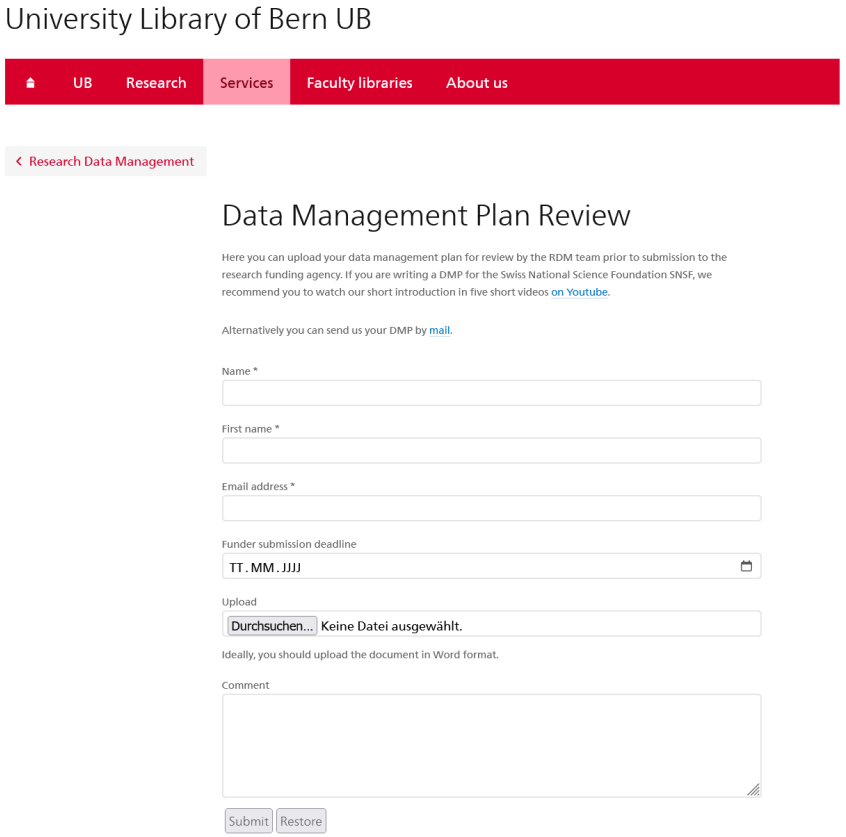
3) The more metadata, the better!

Questions?

Part 3: Support
Data Management Plan Review



Video modules YouTube



Data Management Plan review online
Feedback within 1-3 working days

Support
Training

How to publish & license research data

How to store research data

How to efficiently document research data?

How to plan and update Data Management Plan

How to ethically manage research data

How to collect and organize data and files, formats, names, versions control

@ openscience@unibe.ch

Support
Talks

21 NOVEMBER

17:15 - 18:00 UHR

EVENT
OPEN SCIENCE

Bern Data Talks - Preparing Data for Publication

- Prof. Dr. Tobias Hodel
University of Bern, Digital Humanities

- Prof. Dr. Stefan Brönnimann
University of Bern, Institute of Geography

- Thilo Hirsch
Bern University of Applied Sciences, Bern University of the Arts

Open Science
Newsletter

- Funder news
- Training sessions and courses
- Developments in Open Access and Research Data

To subscribe:

https://www.unibe.ch/ub/osnews

Research Data Management Support
Survey

We would like to learn about your wishes and needs in order to optimize our training and consulting services.
Thank you for participating in this short and anonymous survey!

t1p.de/RDMsurvey

Contact

Dr. Olga Churakova

E-Mail: olga.churakova@unibe.ch

Contact Point:

Medicine, Vetsuisse, Insel

Dr. Gero Schreier

E-Mail: gero.schreier@unibe.ch

Contact Point:

Humanities, Law and Theology

References

Links

http://www.getty.edu/publications/intrometadata/

https://schema.org/

XML-Standards and resources

Open archives initiatives

https://www.openarchives.org/pmh/

References

Metadata

Neiswender C & Montgomery E (2009). *Metadata Interoperability—What Is It, and Why Is It Important?.* In Stocks KI, Neiswender C, Isenor AW, Graybeal J, Galbraith N, Montgomery ET, Alexander P, Watson S, Bermudez L, Gale A, & Hogrefe K. (Eds.), *The MMI Guides: Navigating the World of Marine Metadata, 11-15*.

Riley, J. (2017). *Understanding Metadata: what is metadata, and what is it for?*

Van 't Veer, A. E., & Giner-Sorolla, R. (2016). *Pre-registration in Social Psychology – A discussion and suggested template.* Retrieved from https://osf.io/4frms/

MRI Data: Guidance by the Organization for Human Brain Mapping (OHBM) Committee on Best Practice in Data Analysis and Sharing (COBIDAS)

EEG Data: (German) Empfehlungen zur Erzeugung und Dokumentation von EEG Daten [Recommendations on the Generation and Documentation of EEG data] of the Deutsche Gesellschaft für Klinische Neurophysiologie und Funktionelle Bildgebung.

References

Metadata Standards

Catalogues of metadata standards (browsable by discipline)
https://rdamsc.bath.ac.uk/
https://www.dcc.ac.uk/guidance/standards/metadata

Examples for metadata standards include:
ABCD - Access to Biological Collection Data
AVM - Astronomy Visualization Metadata
CERIF - Common European Research Information Format

(Examples, continued)
CF (Climate and Forecast) Metadata Conventions
CIF - Crystallographic Information Framework
Darwin Core
DataCite Metadata Schema
DDI - Data Documentation Initiative
Dublin Core
EML - Ecological Metadata Language
FGDC/CSDGM - Federal Geographic Data Committee Content Standard for Digital Geospatial Metadata

References

Links

Controlled vocabularies
https://guides.lib.unc.edu/metadata/controlled-vocab
EU Vocabularies
ReadMe file (Cornell University)
Readme_Template_EN_20220307(1).txt (3KB)
Clinical Metadata Standard

References

Documentation Tools

LIMS Open BIS https://openbis.ch/

ELN vs. LIMS: How to Make a Choice? - YouTube

SLIMS for analitical labs, pharma https://explore.agilent.com/AgilentSlims

Note-Taking Software
Examples

- Microsoft OneNote → Guide to using OneNote as a Research Notebook
- Evernote – note taking, organising, task management and archiving. Imports and exports in Evernote XML
- Notion – Wiki, docs & notes
- Transkribus - to help you transcribing documents. This is a comprehensive platform for the digitization, AI-powered text recognition, transcription and searching of historical documents – from any place, any time, and in any language.

Open Source Applications

Cross-Platform

TiddlyWiki

Standard note, Boostnote, Turtl, client-side encryption

Joplin → end-to-end encryption

Benchling

This is a free cloud-based ELN platform that also provides Molecular Biology tools. Free version available (academia only) with 10 GB of storage space.

Zim

Desktop wiki application, open source, Linux, Windows

Thank you

for your attention

Open Science Team
openscience@unibe.ch