Data documentation through metadata

GEO 802 Fall 2020, Data Information Literacy

Anna C. Véron, Dr. sc. nat.

Lesson 6: Data documentation through metadata

→ Definition of metadata

→ Why we need metadata

→ FAIR data

→ Metadata standards

→ How to write quality metadata

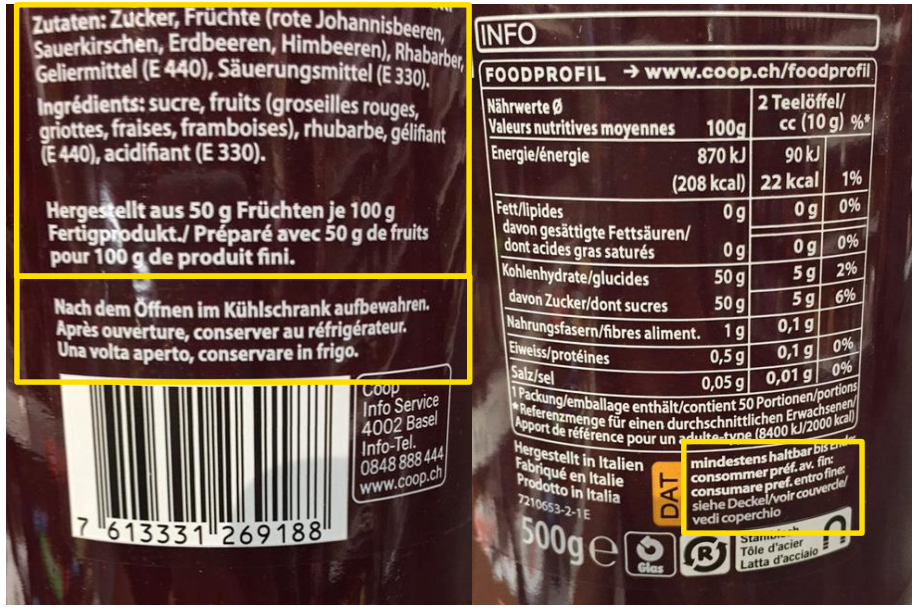
What is metadata?

Data about data

(Almost) The same product – two different qualities of metadata.

https://fddb.info/db/de/lebensmittel/selbst_gemacht_erdbeermarmelade_mit_gelierzucker_2_plus_1/foto.html#201444

https://www.foodrepo.org/ch/products/3028



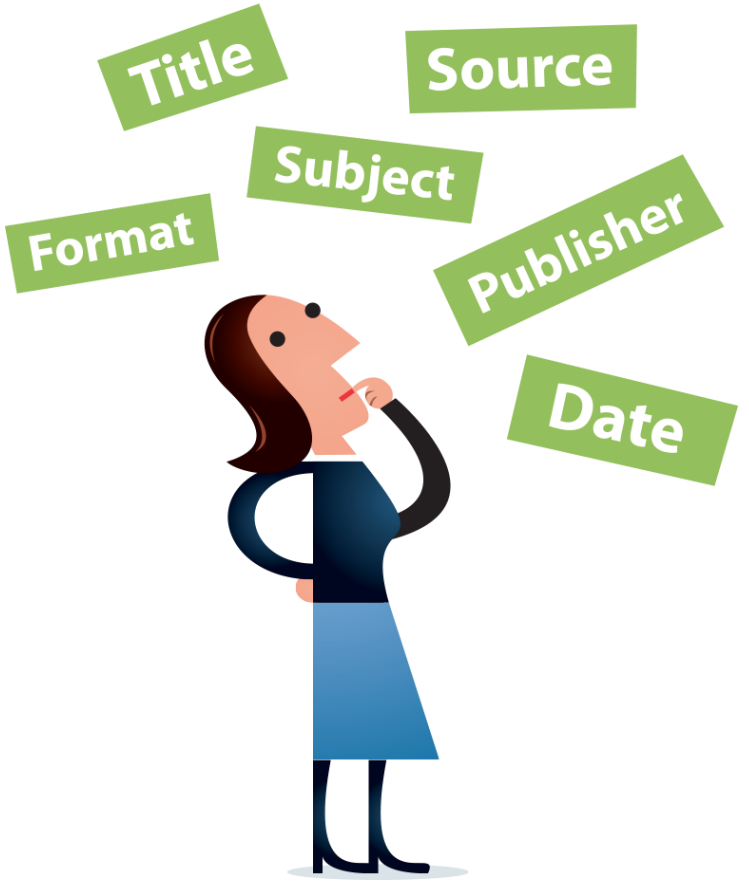
What is metadata?

- "Structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource." *NISO, Understanding Metadata*

- It can be used to describe physical items as well as digital items (documents, audio-visual files, images, datasets, etc.)

- Metadata can take many different forms, from free text (such as read-me files) to standardized, structured, machine-readable content

- For data to be useful, it will also need subject-specific metadata (reagent names, experimental conditions, population demographic…)



What is metadata?

Metadata is «Data reporting»

→ **WHO** created the data?

→ **WHAT** is the content of the data?

→ **WHEN** were the data created?

→ **WHERE** is it geographically?

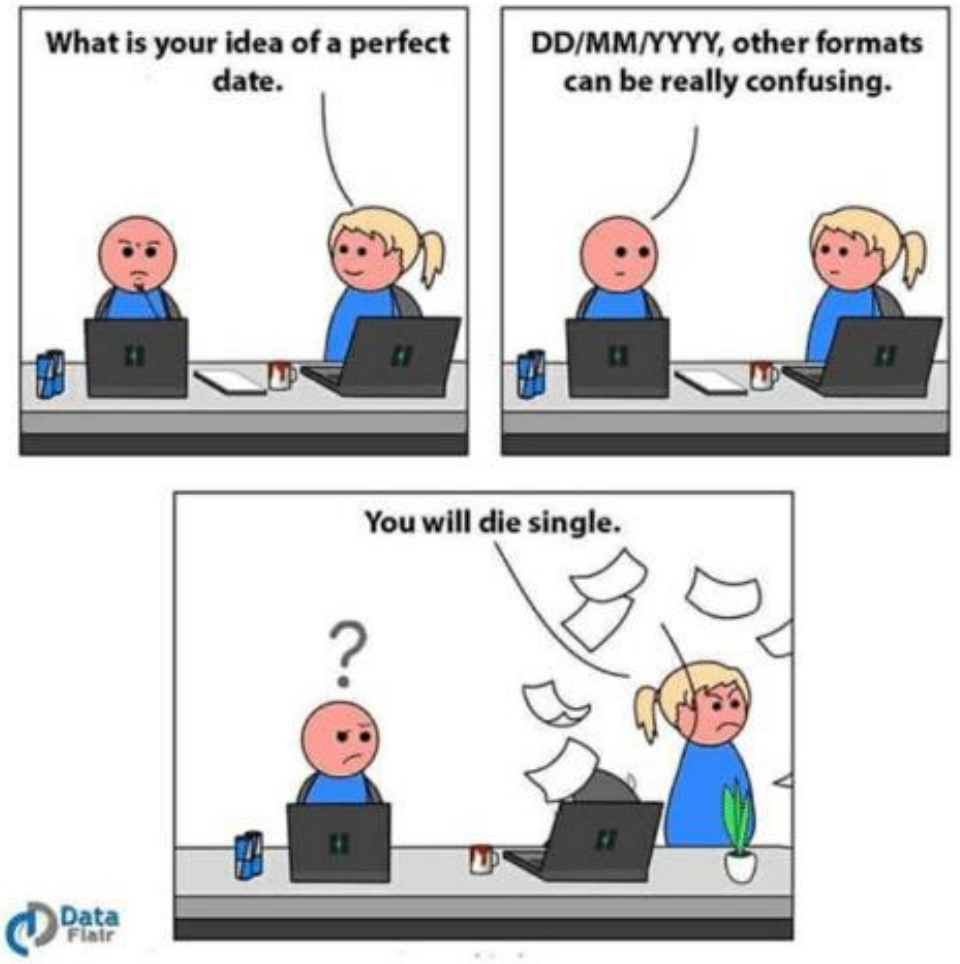
→ **HOW** were the data developed?

→ **WHY** were the data developed?

Working with data

When you **receive a dataset** from an external source, what types of details do you want to know about the data?

When you **provide data** to someone else, what types of information should you include with the data?



Working with data

- **Receiving data**:
  - What are the data gaps?
  - What processes were used for creating the data?
  - Are there any fees associated with the data?
  - In what scale were the data created?
  - What do the values in the tables mean?
  - What software do I need in order to read the data?
  - What projection are the data in?
  - Can I give these data to someone else?

- **Providing data**:
  - Why were the data created?
  - What limitations, if any, do the data have?
  - What does the data mean?
  - How should the data be cited if it is re-used in a new study?

Lesson 6: Data documentation through metadata

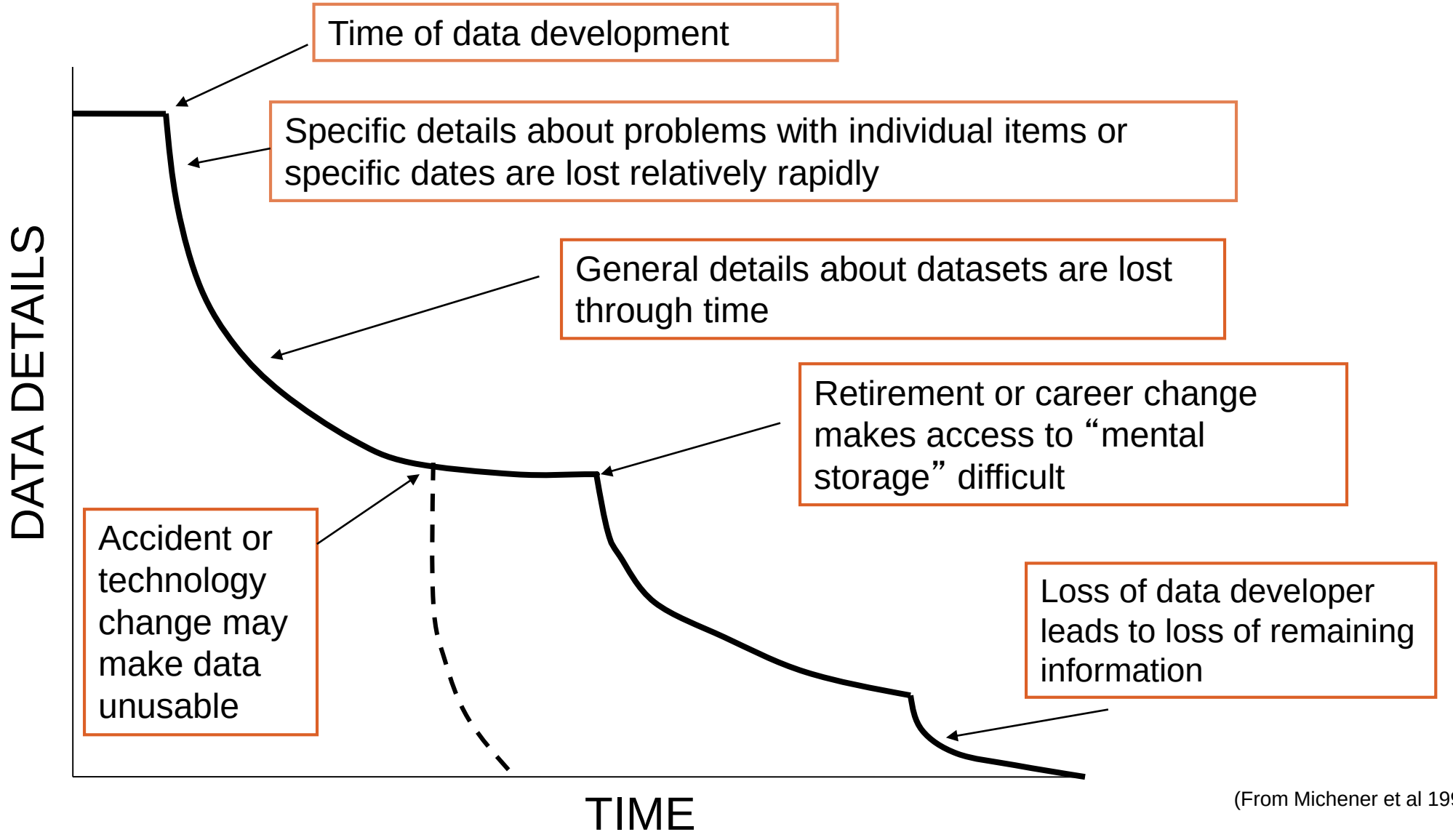
- ✓ Definition of metadata
- → Why we need metadata
- → FAIR data
- → Metadata standards
- → How to write quality metadata

Why we need metadata

"The metadata accompanying your data should be written for a user ***20 years into the future*** - what does that person need to know to use your data properly? Prepare the metadata for a user who is unfamiliar with your project, methods, or observations."

Oak Ridge National Laboratory Distributed Active Archive Center for Biogeochemical Dynamics (ORNL DAAC)

When metadata are bad… «Information decay» happens faster



(From Michener et al 1997)

The value of metadata

Metadata allows data creators to…

- Avoid data duplication
- Share reliable information
- Publish data and receive citations
  → promotes a scientist's work and their contributions to their field

Metadata gives data users the ability to…

- Search, retrieve, and evaluate data set information from both inside and outside an organization
- Find data: Determine what data exists for a geographic location and/or topic
- Determine applicability: Decide if a data set meets a particular need
- Discover how to acquire the dataset you identified; process and use the dataset

Lesson 6: Data documentation through metadata

- ✓ Definition of metadata
- ✓ Why we need metadata
- → FAIR data
- → Metadata standards
- → How to write quality metadata

FAIR principles

- Introduced in 2016 by FORCE 11
  (= representatives from science, funding institutions, publishers, libraries, archives)
- Goal: optimal processing of research data for both human and machine
- 15 Principles
- Explanation by the SNF:
  https://tinyurl.com/SNFfair

Findability

- **Persistent identifier (PID): e.g. Digital Object Identifier (DOI)**

- Descriptive metadata in a machine readable format
  - Title, author / creator of data
  - Context, quality, condition and characterization of the data
  - How was the data generated?
  - Which information is needed to interpretate the data?

Data that really saves lives (and possibly your organisation) by Fredric Landqvist. findwise.com/blog

Accessibility

– Open access to anyone in the world with a computer and internet connection (no charge, no other access restrictions)

Limitations

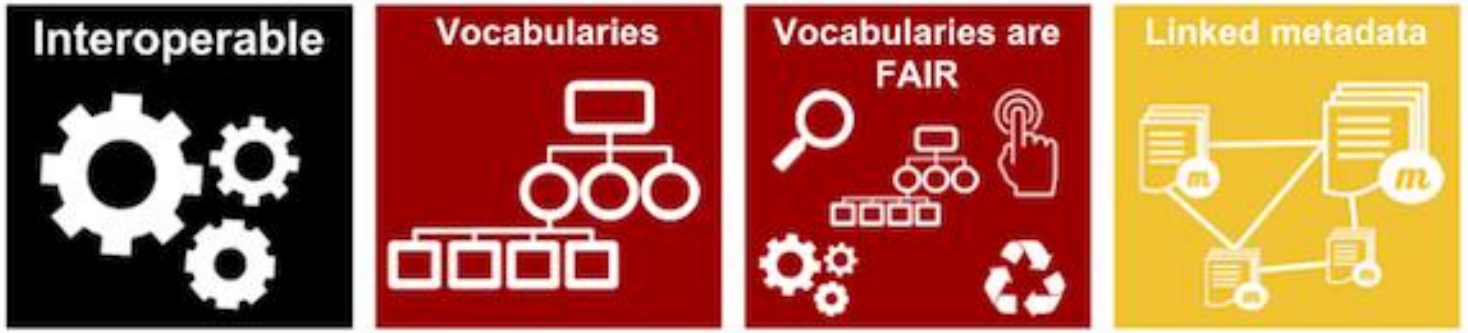
- Data which are subject to data protection and privacy laws (e.g. involving living individuals)
- Data from international collaboration with countries that have laws prohibiting the open sharing of data

– At least the **metadata have to be accessible**!

Data that really saves lives (and possibly your organisation) by Fredric Landqvist. findwise.com/blog

Interoperability

- Data and metadata have to be fully compatible between different computer operating systems
- **Open file formats** (files can be used with with freely available software)
- Use of **controlled vocabulary** with an easily findable and accessible documentation
- Citation of relevant / associated data sets



Data that really saves lives (and possibly your organisation) by Fredric Landqvist. findwise.com/blog

Re-usability

- Metadata must contain any **information necessary to properly understand and use the data**. The categories of metadata must be explained or self-explanatory.
- Data needs to be **reliable (reproducible)** and **understandable**!
- Include information about the **license** in the metadata. Whenever possible, the data must be **labelled for reuse**.

Data that really saves lives (and possibly your organisation) by Fredric Landqvist. findwise.com/blog

Exercise 6.1: FAIR Data

- Compare the metadata of two different datasets
- Are the FAIR principles implemented? If yes, how? What is missing?

https://tinyurl.com/Zenodo

PANGAEA.
Data Publisher for Earth & Environmental Science

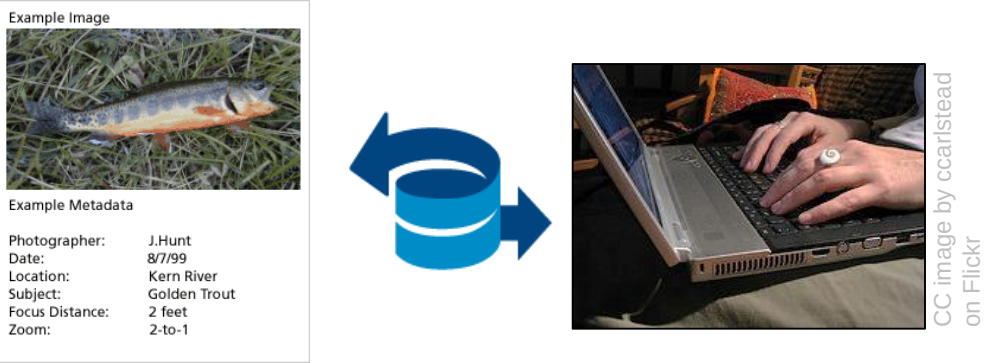
https://tinyurl.com/panga123

Lesson 6: Data documentation through metadata

- ✓ Definition of metadata
- ✓ Why we need metadata
- ✓ FAIR data
- → Metadata standards
- → How to write quality metadata

What is a metadata standard?

- **A Standard provides a structure to describe data with**
  - Common terms to allow consistency between records
  - Common definitions for easier interpretation
  - Common language for ease of communication
  - Common structure to quickly locate information
- **In search and retrieval, standards provide:**
  - Documentation structure in a reliable and predictable format for computer interpretation
  - A uniform summary description of the dataset



CC image by ccarlstead on Flickr

Examples of metadata standards

- **Darwin Core** | biological diversity, taxonomy
- **Dublin Core |** general
- **DDI** (Data Documentation Initiative) | social & behavioral sci.
- **DIF** (Directory Interchange Format) | environmental sci.
- **EML** (Ecological Metadata Language) | ecology, biology
- **ISO** 19115| geographic data

Examples of metadata standards

- **Dublin Core Element Set**
  - Emphasis on web resources, publications
  - http://dublincore.org/documents/dces/

- **FGDC Content Standard for Digital Geospatial Metadata (CSDGM)**
  - Emphasis on geospatial data
  - http://www.fgdc.gov/metadata/geospatial-metadata-standards

- **Biological Data Profile (BDP) of the CSDGM**
  - Profile to the CSDGM emphasis on biological data (and geospatial)
  - https://www.fgdc.gov/standards/projects/metadata/biometadata/index_html

- **ISO 19115/19139 Geographic information: Metadata**
  - Emphasis on geospatial data and services
  - http://www.fgdc.gov/metadata/geospatial-metadata-standards#fgdcendorsedisostandards

Examples of metadata standards

- **Ecological Metadata Language (EML)**
  - Focus on ecological data
  - http://knb.ecoinformatics.org/eml_metadata_guide.html

- **Darwin Core**
  - Emphasis on museum specimens
  - http://rs.tdwg.org/dwc/index.htm

- **Geography Markup Language (GML)**
  - Emphasis on geographic features (roads, highways, bridges)
  - http://www.opengeospatial.org/standards/gml

- **OGC® WaterML**
  - WaterML 2.0 is a standard information model for the representation of water observations data
  - http://www.opengeospatial.org/standards/waterml

Exercise 6.2: Metadata Standards

- Browse through metadata standards by discipline.
- Take note of standards that might be relevant for your field.

→ http://www.dcc.ac.uk/resources/metadata-standards

→ http://rd-alliance.github.io/metadata-directory/tools/

Lesson 6: Data documentation through metadata

- ✓ Definition of metadata
- ✓ Why we need metadata
- ✓ FAIR data
- ✓ Metadata standards
- → How to write quality metadata

Steps to create quality metadata

- Work with a data management plan! (Lesson 10)
- **Document your research process and create metadata while you are creating and analyzing data!**

- Organize your information
  - Did you use a lab notebook or other notes during the data development process that define measurements and other parameters?
  - Do you have the contact information for colleagues you worked with?
  - What about citations for other data sources you used in your project?

- Have someone else read your record
- Revise the record, based on comments from your reviewer
- Review once more before you publish

Tipps for writing quality metadata

- **Do not use jargon**
- **Define technical terms and acronyms:**
  - CA, LA, GPS, GIS : what do these mean?
- **Clearly state data limitations**
  - E.g., data set omissions, completeness of data
  - Express considerations for appropriate re-use of the data
- **Use “none” or “unknown” meaningfully**
  - “None” usually means that you knew about data and nothing existed (e.g., a “0” cubic feet per second discharge value)
  - "Unknown” means that you don’t know whether that data existed or not (e.g., a null value)

Tipps for writing quality metadata

- **Select keywords wisely**
- **Use descriptive and clear writing**
- **Fully qualify geographic locations**
- **Use thesauri (controlled vocabulary) for keywords whenever possible**

Controlled Vocabulary / Thesaurus

- An organized arrangement of terms and phrases
- Used to index content and/or to retrieve content through browsing or searching
- Controlling your Language: a Directory of Metadata Vocabularies by JISC (UK)

- **E.g. hierarchical list of minerals (GeoRef)**

oxysulfides
phosphates
phosphides
selenates
selenides
selenites
silicates (use a narrower term below if dealing with
specific mineral; otherwise larger group)
aluminosilicates
orthosilicates
sorosilicates
orthosilicates, axinite group
orthosilicates, chevkinite group
orthosilicates, epidote group
orthosilicates, melilite group
orthosilicates, pumpellyite group
orthosilicates, thortveitite group
nesosilicates

Summary of Lesson 6

Metadata is documentation of data.

Metadata allows data to be discovered, accessed, and re-used.

A metadata standard provides structure and consistency to data documentation.

Document your process while you are creating and analyzing data.

Metadata completes a dataset.

Creating quality metadata is in your OWN best interest!