Organizing, Describing & Processing Data

J.S. Love

Data Steward, Industrial Design Engineering

TU Delft

TU Delft

Today's Agenda

- Files Names & Formats
- Data Organization: Folder structure
- Versioning
- Documenting Data & Metadata

Today's Watchword:

Transparency

What kinds of DATA do you have?

Examples of Research Data

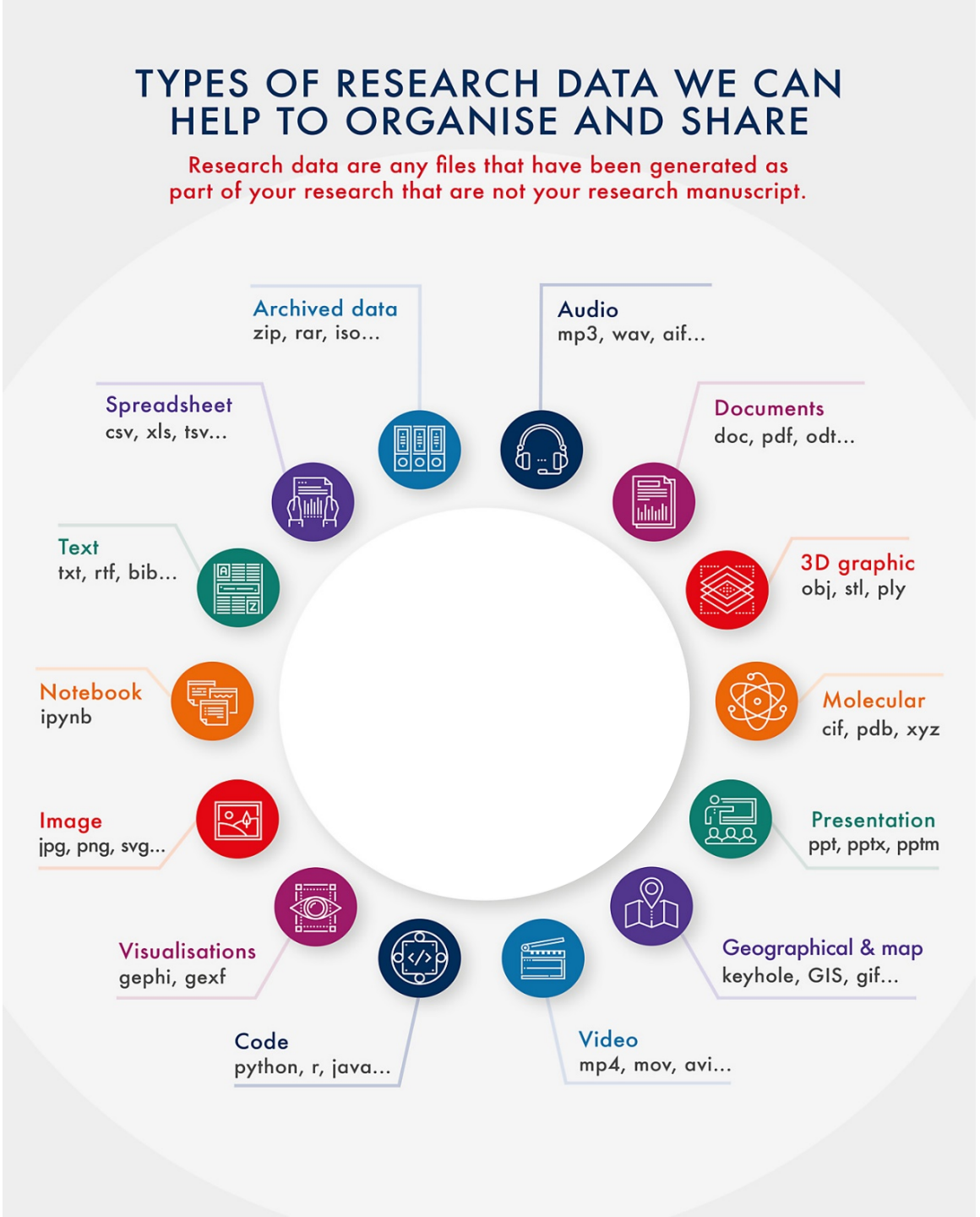


Image Adapted from: Admin, Springer Nature; Astell, Mathias (CC-BY)

File names & formats - Which is better?

20141020145256673.pdf	2014110615475230 9.pdf	FUNC_04&VLN_1102&DEV_0011&SUBSYS_10280708.reg
IEEEMeetingNotes_20170304.pdf	AlgaeColonyObservations1964-1993.txt	Hawking2003Excerpts.rtf
Diss Chapter 1, new.docx	Appendix 2.docx	Bibliography.docx

File Names & Formats – Good Practice

File Naming:

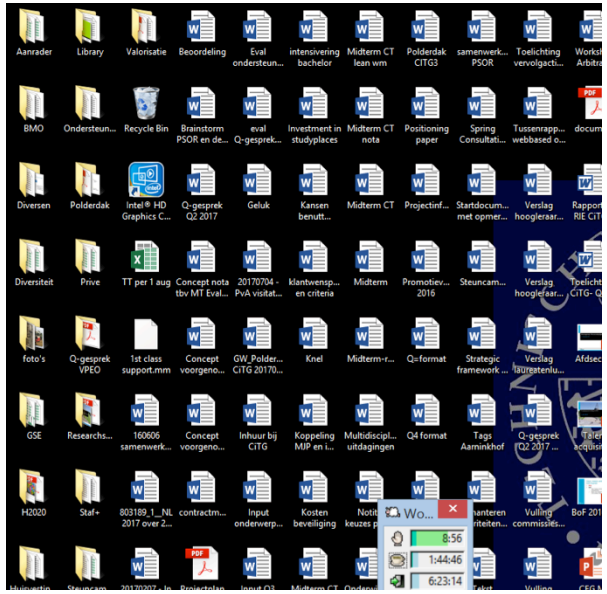
- Be descriptive
- Use: dates (yyyymmdd), underscores, hyphens
- Avoid!: spaces, question marks, commas
- Change names immediately, not later

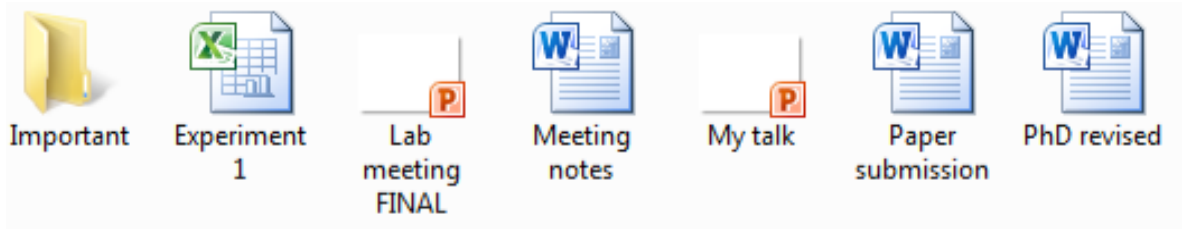
File Formats

- Open (e.g. PDF/A, .txt, .csv) > proprietary (.docx, .spv)
- Resources:
  - Library of Congress Recommended Formats: https://www.loc.gov/preservation/resources/rfs/
  - Digital Preservation Coalition Handbook: https://www.dpconline.org/handbook/technical-solutions-and-tools/file-formats-and-standards
  - UK Data Service Recommended Formats: https://www.ukdataservice.ac.uk/manage-data/format/recommended-formats

General Tips from MANTRA: https://mantra.edina.ac.uk/organisingdata/

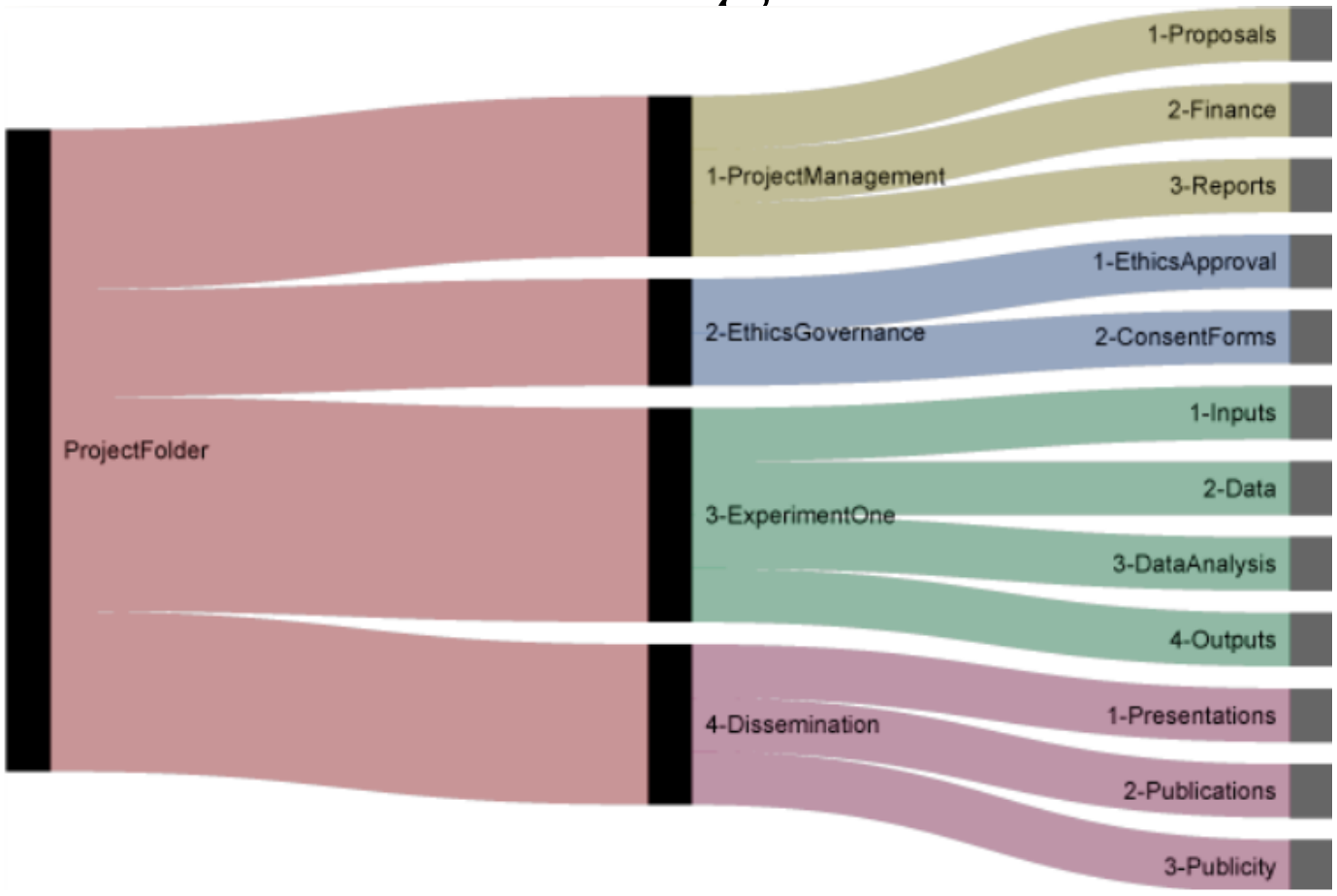
Data Structuring – how it often looks





Slide Credit: Maria Cruz (VU) & Yasemin Türkyilmaz-van der Velden (TUD)

Structuring Folders



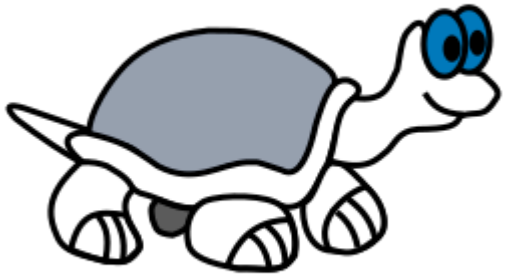
Research Project Folder Structure from Nikola Vukovic:
http://nikola.me/folder_structure.html

Resources and examples of file & folder management

Organizing files and folders from the Wageningen Data Competence Center: https://www.wur.nl/en/Value-Creation-Cooperation/WDCC/Data-Management-WDCC/Doing/Organising-files-and-folders.htm

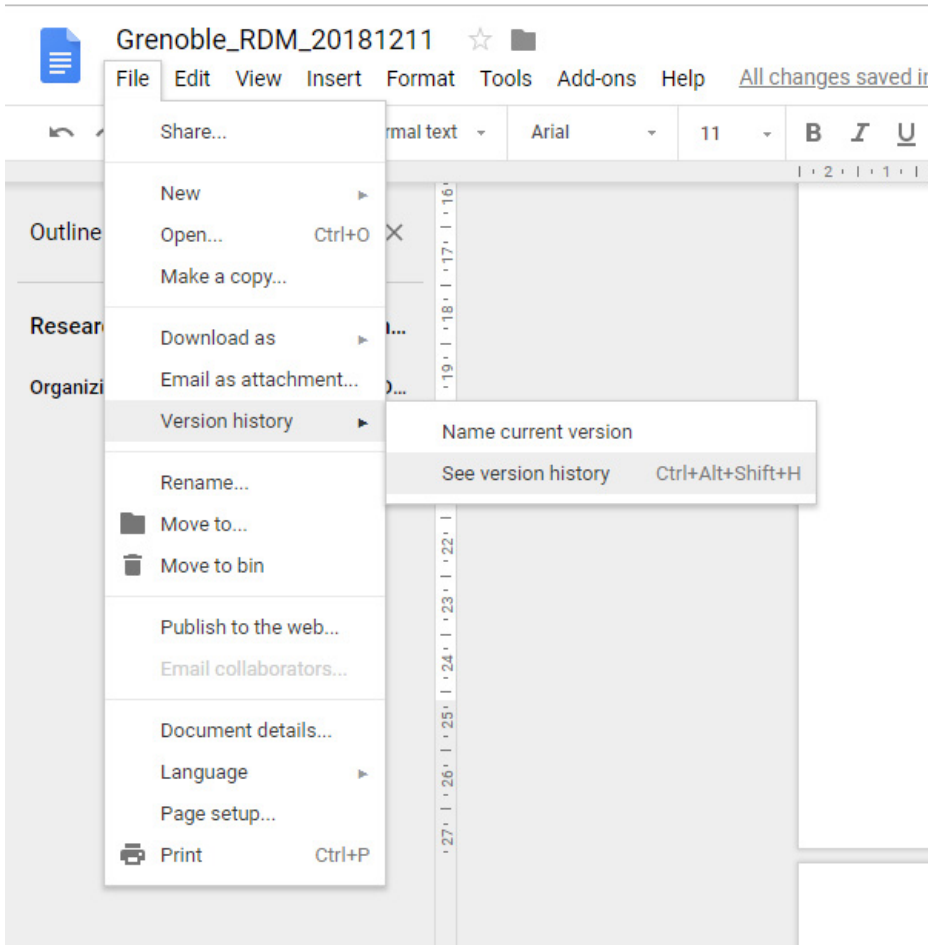
Versioning – What is it?

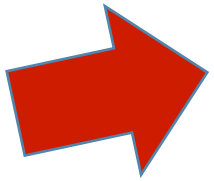
Versioning – it’s good for you

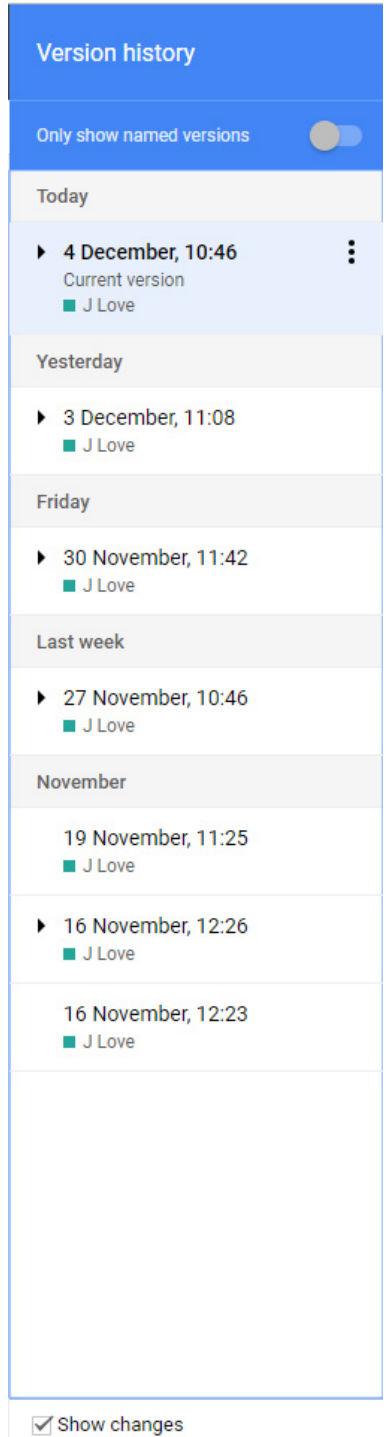


TortoiseSVN

Versioning lite: GoogleDocs, Dropbox, etc.







Versioning in GoogleDocs

Resources for Versioning

Intro articles to versioning:

- Git: https://www.makeuseof.com/tag/file-versioning-git/
- GoogleDocs, Dropbox, MS Word and Draft: https://www.makeuseof.com/tag/not-just-for-coders-top-version-control-systems-for-writers/

An Introduction to Version Control using GitHub Desktop by Daniel van Strien on The Programming Historian: https://programminghistorian.org/en/lessons/getting-started-with-github-desktop

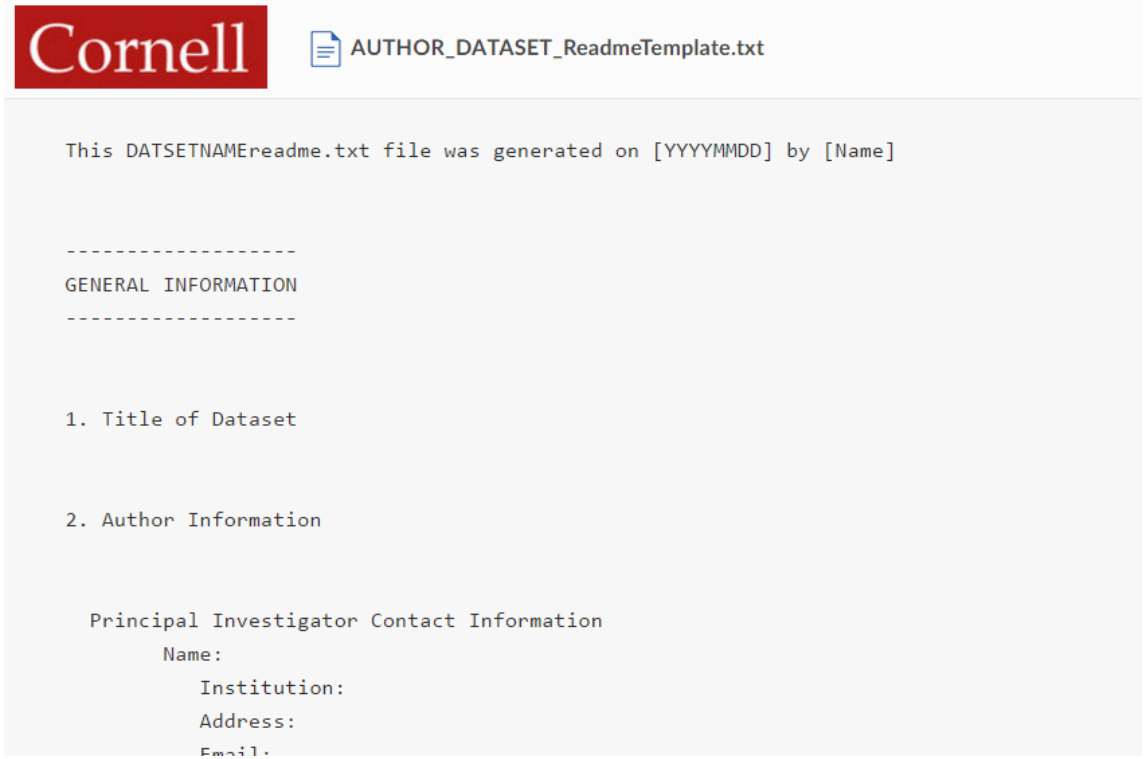
Documentation

- The glories of the readme file
- Ontologies & standards
- Metadata

Image credit: Rodhullandemu CC-BY-SA

Many ways of describing data

How to create useful README files :
https://data.research.cornell.edu/content/readme


Cornell

AUTHOR_DATASET_ReadmeTemplate.txt

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


-------------------
GENERAL INFORMATION
-------------------


1. Title of Dataset


2. Author Information


  Principal Investigator Contact Information
        Name:
           Institution:
           Address:
           Email:
```

README files template: https://cornell.app.box.com/v/ReadmeTemplate

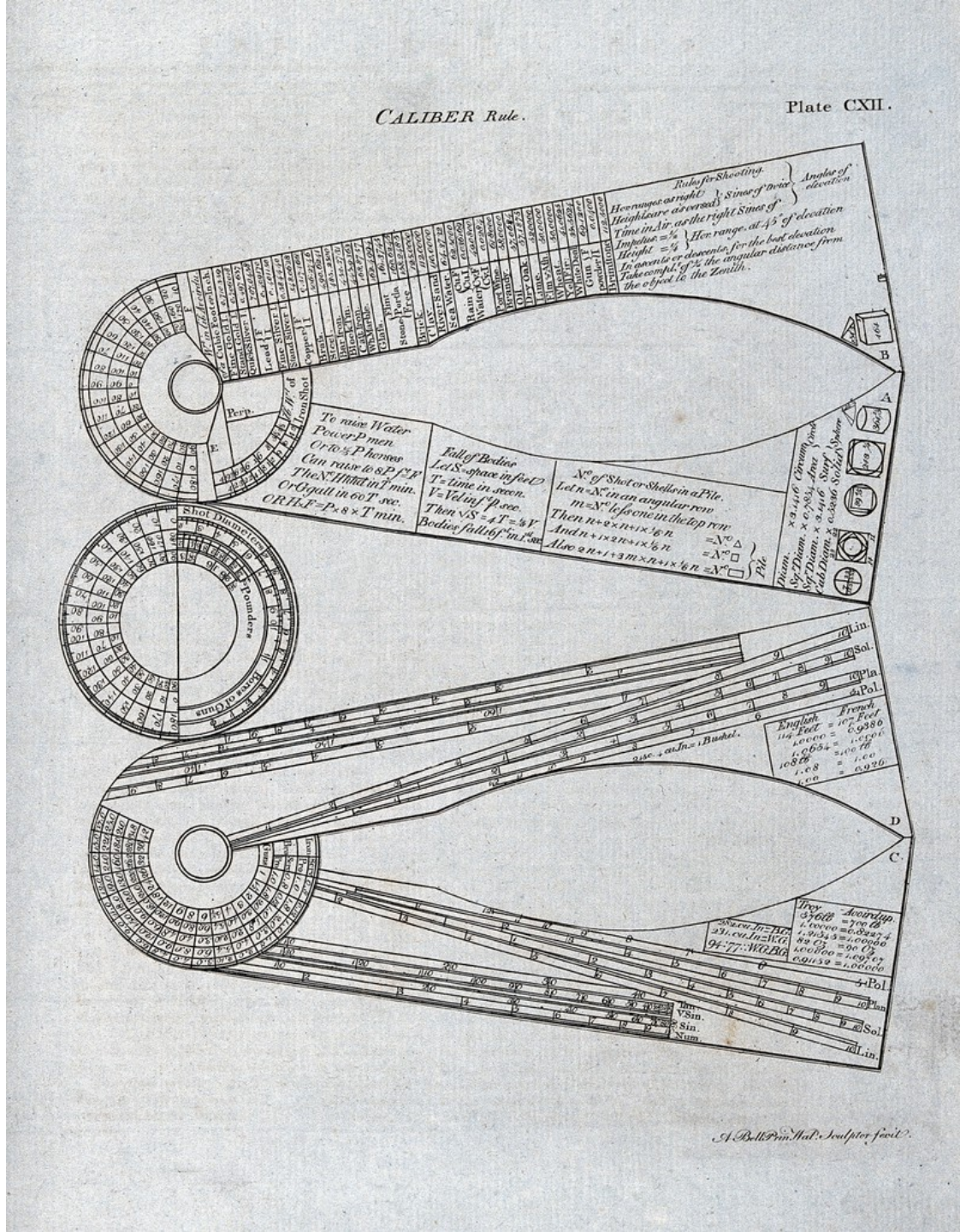
Slide credit: Marta Teperek

Standards & Ontologies

- Discipline-specific
- Find in major papers of your field and ask colleagues

Still looking for a standard vocabulary in your field? Try here:

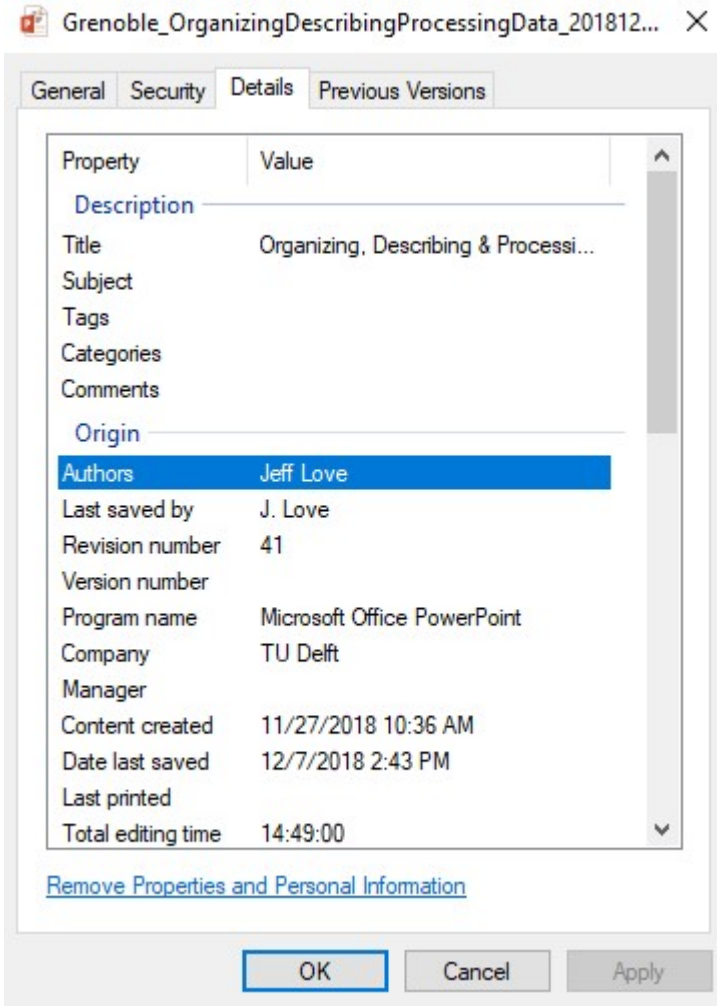
- https://fairsharing.org/standards/
- https://lov.linkeddata.es/dataset/lov/vocabs



'Calipers for artillery measurement, with many formulae engraved on them.' Wellcome Collection (CC-BY)

What is Metadata?

Metadata – data about data



```
−<gmi:MI_Metadata xsi:schemaLocation="http://www.isotc211.org/2005/gmi https://ngdc.noaa.gov/metadata/published/xsd/schema.xsd">
  −<gmd:fileIdentifier>
      <gco:CharacterString>gov.noaa.ngdc.mgg.multibeam:KN166L14_Multibeam</gco:CharacterString>
    </gmd:fileIdentifier>
  −<gmd:language>
      <gco:CharacterString>eng; USA</gco:CharacterString>
    </gmd:language>
  −<gmd:characterSet>
      <gmd:MD_CharacterSetCode codeListValue="utf8" codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.xml#MD_CharacterSetCode">utf8</gmd:MD_CharacterSetCode>
    </gmd:characterSet>
  −<gmd:hierarchyLevel>
      <gmd:MD_ScopeCode codeListValue="series" codeList="https://www.ngdc.noaa.gov/metadata/published/xsd/schema/resources/Codelist/gmxCodelists.xml#MD_ScopeCode">series</gmd:MD_ScopeCode>
    </gmd:hierarchyLevel>
  −<gmd:hierarchyLevelName>
      <gco:CharacterString>Multibeam Collection Level Metadata</gco:CharacterString>
    </gmd:hierarchyLevelName>
  −<gmd:contact xlink:title="Evan Robertson">
    −<gmd:CI_ResponsibleParty xsi:schemaLocation="http://www.isotc211.org/2005/gmi https://www.ngdc.noaa.gov/metadata/published/xsd/schema.xsd" uuid="af436120-c4a6-11e2-8b8b-0800200c9a66">
      −<gmd:organisationName>
        −<gco:CharacterString>
            DOC/NOAA/NESDIS/NCEI > National Centers for Environmental Information, NESDIS, NOAA, U.S. Department of Commerce
          </gco:CharacterString>
        </gmd:organisationName>
      −<gmd:positionName>
          <gco:CharacterString>Multibeam Bathymetry Data Manager</gco:CharacterString>
        </gmd:positionName>
      −<gmd:contactInfo>
        −<gmd:CI_Contact>
          −<gmd:phone>
            −<gmd:CI_Telephone>
              −<gmd:voice>
                  <gco:CharacterString>303-497-5414</gco:CharacterString>
                </gmd:voice>
              </gmd:CI_Telephone>
            </gmd:phone>
          −<gmd:address>
            −<gmd:CI_Address>
              −<gmd:deliveryPoint>
                  <gco:CharacterString>325 Broadway E/NE42</gco:CharacterString>
                </gmd:deliveryPoint>
              −<gmd:city>
                  <gco:CharacterString>Boulder</gco:CharacterString>
                </gmd:city>
              −<gmd:administrativeArea>
```


Metadata Standards

General

Dublin core

MODS

METS

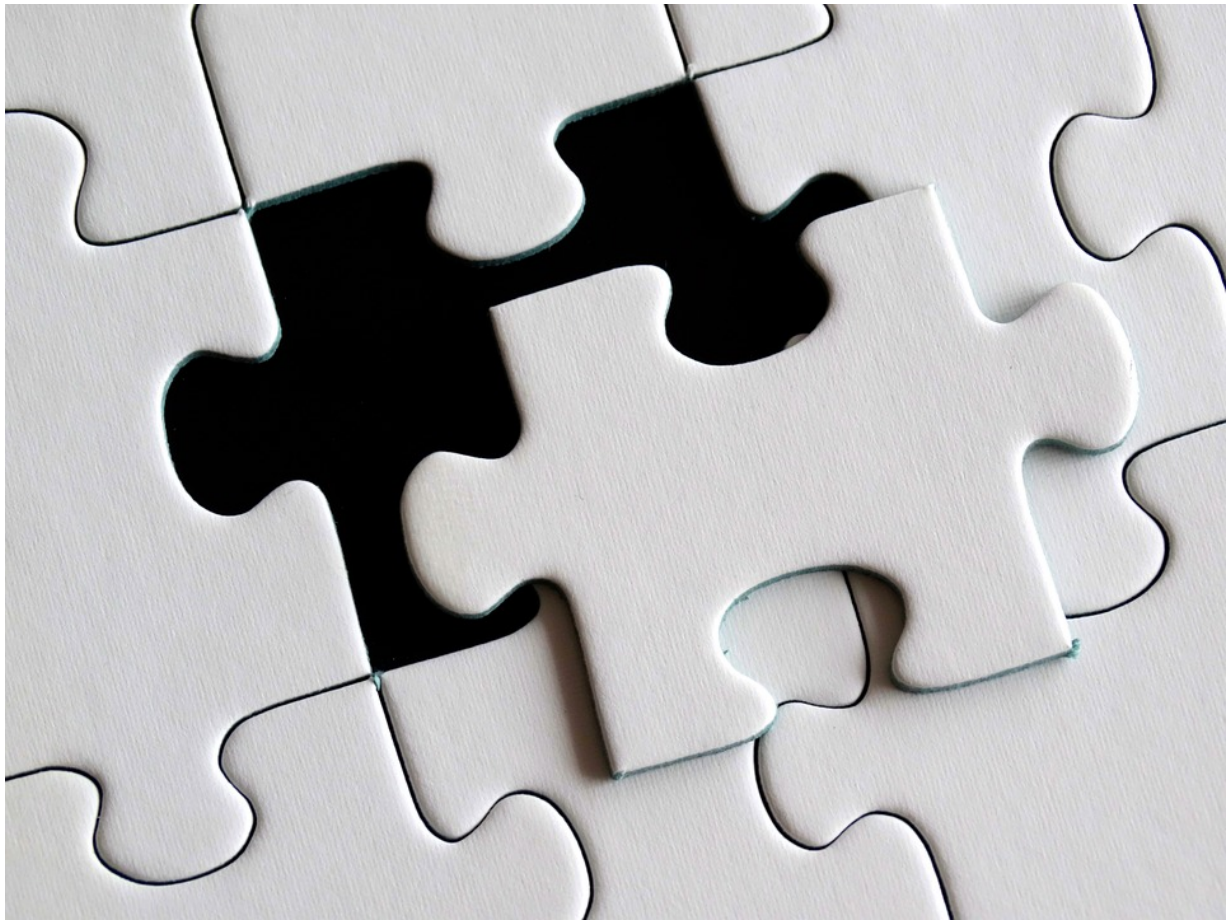
Discipline-specific

GMD

DDI

Multibeam collection for KN166L14: Multibeam data collected aboard Knorr from 2002-08-03 to 2002-09-09, Reykjavik, Iceland to Reykjavik, Iceland (https://data.nodc.noaa.gov/cgi-bin/iso?id=gov.noaa.ngdc.mgg.multibeam:KN166L14_Multibeam)

Metadata mix & match



Metadata mix & match

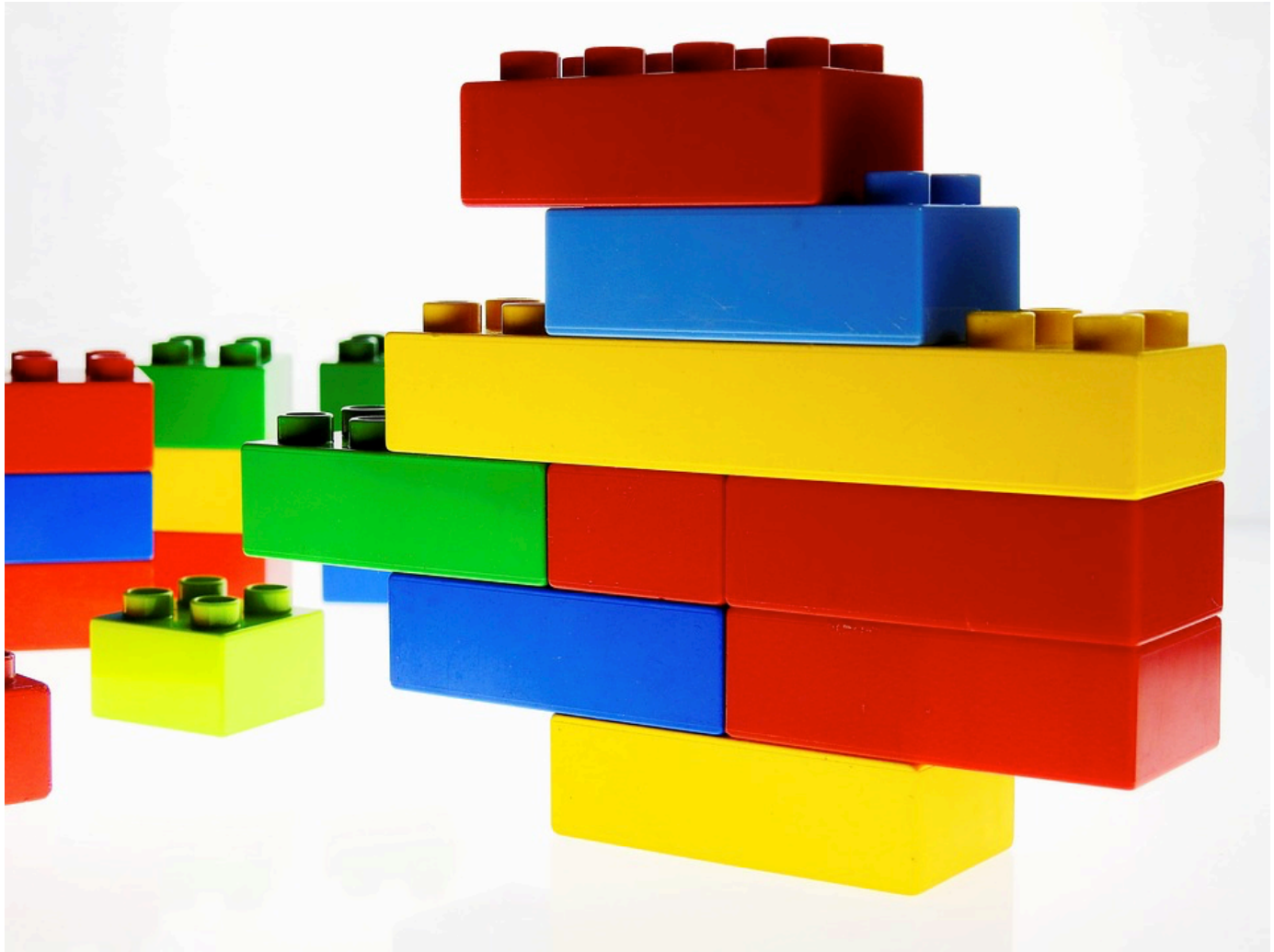
https://data.nodc.noaa.gov/cgi-bin/iso?id=gov.noaa.ngdc.mgg.multibeam:KN166L14_Multibeam

Exercise – Documentation & Metadata

Directions:

1. Use your set of pieces to construct an object (take a picture if you want)
2. Write down instructions on how to build that object
3. Pull apart your object

In five minutes, you'll pass on your directions and pieces to another group, who will try to rebuild the object you made.



Parting Words

Exploit Research Infrastructures, such as:

- The European Marine Biological Resource Centre (EMBRC): http://www.embrc.eu/
- Survey of Health, Ageing and Retirement (SHARE): http://www.share-project.org/
- The Research Data Alliance (RDA): https://www.rd-alliance.org/

Learn to script/code! (HTML, Python, PHP, etc.)

W3schools (HTML, CSS, SQL, PHP, ASP, JS): https://www.w3schools.com/

Data Carpentry: http://www.datacarpentry.org/lessons/