

# Research Data and Data Management Planning

Markus Stocker

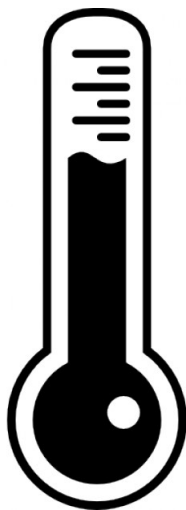
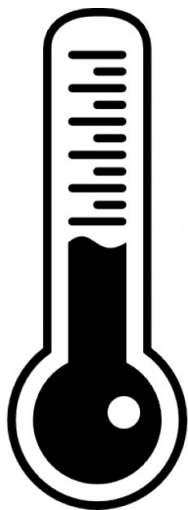
September 12, 2017

# Outline

- What are research data
- Research data lifecycle
- Data types, formats, models and standards
- Metadata
- Data management, plans and planning tools



Datum is ultimately reducible to a lack of uniformity [1]



# Define data

- Entities, physical or digital, used as evidence of phenomena [2]
- A reinterpretable representation of information [3]
- There is no consensus definition
- Even institutions that curate data may not define what they curate

# Data examples

- Not just spreadsheets of numbers, also
- Sequences of bits
- Characters on a page
- Recording of sounds
- Physical and biological specimens
- Images
- Software

# Define research data

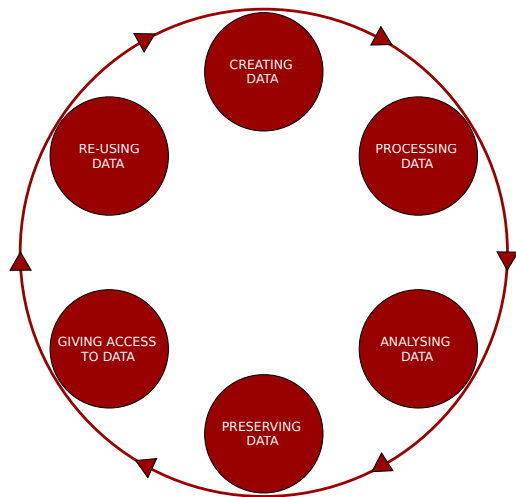
- Unsurprisingly, there is no consensus on the definition
- Factual material [...] necessary to validate research findings [4]
- Everything needed to reproduce a given scientific output [5]



# Research data examples

- In addition to the obvious, e.g. data files
- Notebooks, e.g. laboratory, field, diaries, ...
- Questionnaires, audio and video tapes
- Models and scripts
- Workflows and protocols

# Research data lifecycle



Adapted from <http://www.data-archive.ac.uk/create-manage/life-cycle>

# Research data types

- Observational data
  - ▶ Result from recognizing, noting or recording facts
  - ▶ Collected by human observation, surveys, instruments
  - ▶ Typically difficult or impossible to reproduce
  - ▶ Example: Ocean temperature measurements at spacetime locations
- Experimental data
  - ▶ Result of procedures in controlled conditions
  - ▶ In theory reproducible but may be expensive
  - ▶ Example: Results of chemical analysis in a laboratory
- Computational data
  - ▶ Result in executing computer models, simulations, or workflows
  - ▶ Reproducible if software and input available
  - ▶ Example: Output of a plant disease pressure model

# Data formats

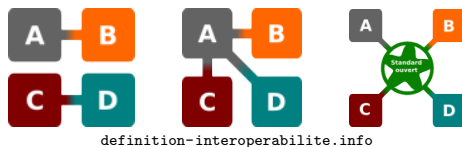
- Serialized representation of data
- Comma separated values (CSV) is a common format
- JPEG, TIFF, PNG, GIF for raster images
- SVG, EPS, PDF for vector graphics
- NetCDF, HDF5 for array scientific data

# Data models

- Abstract formalization of objects and relationships in a domain
  - ▶ There are lakes, rivers, and mountains
  - ▶ Lakes and rivers have a depth
  - ▶ Mountains have an elevation
  - ▶ Rivers have a length
- Set of concepts used to define formalizations
  - ▶ Entity-relationship data model
  - ▶ Graph data model
  - ▶ Geographic data model

# Standards

- Heterogeneity in formats and models hinders interoperability
- Characteristic of system to work with other systems



- Syntactic and semantic interoperability
- If you can, use (de facto) standards, recommendations
- Look for wide acceptance in your community

# Metadata

- Metadata is data about other data
- Metadata describes, explains, locates data
- Supports discovery, retrieval, use, management of data
- May be created manually or automatically
- What is data for someone may be metadata for someone else
- Examples
  - ▶ Data about observational data, e.g. about sensor and property
  - ▶ Data about published data, e.g. title, authors, identifiers
  - ▶ Phone call content vs. phone number, call duration

# Data management

- Organization, storage, preservation, and sharing of data
- Note: Data *and* metadata
- From issues such as consistent file naming conventions
- To safeguarding access over the next 10 years



# Why data management

- Saves time and resources in the long run
- Prevents errors and increases quality of research
- Supports replication and validation of results
- Increasingly required by funders and publishers

# Planning data management

- Good data management needs to be carefully planned
- Write a Data Management Plan
- See exercise

# Online training material

- ESIP Federation Data Management Short Course for Scientists
  - ▶ <http://commons.esipfed.org/datamanagementshortcourse>
- University of Edinburgh, Research Data Management Training
  - ▶ <http://mantra.edina.ac.uk/>
- Coursera, Research Data Management and Sharing
  - ▶ <https://www.coursera.org/learn/data-management>
  - ▶ Starts September 11

## Take aways

# References

- [1] Luciano Floridi. *The Philosophy of Information*. Oxford University Press, 2011. ISBN 978-0-19-923239-0.
- [2] Christine L. Borgman. *Big Data, Little Data, No Data: Scholarship in the Networked World*. The MIT Press, 2015. ISBN 9780262028561.
- [3] CCSDS. Reference Model for an Open Archival Information System (OAIS). Recommended Practice CCSDS 650.0-M-2, The Consultative Committee for Space Data Systems, Washington, DC, USA, June 2012. URL <https://public.ccsds.org/Pubs/650x0m2.pdf>.
- [4] EPSRC. Research Data. URL <https://www.epsrc.ac.uk/about/standards/researchdata/scope/>.
- [5] Alisa Surkis and Kevin Read. Research data management. *Journal of the Medical Library Association : JMLA*, 103(3): 154–156, jul 2015. doi: 10.3163/1536-5050.103.3.011. URL <https://doi.org/10.3163/1536-5050.103.3.011>.

Slide 3: Joan Miró (1968). Landscape. Acrylic on canvas. Fundació Joan Miró, Barcelona.

<https://www.fmirobcn.org/en/colection/catalog-works/5442/p-landscape-p>

Slide 5: On defining data, see <http://pitt.libguides.com/managedata>

Slide 14: On research data management, see <http://pitt.libguides.com/managedata>