How FAIR is your data?

An introduction to research data management

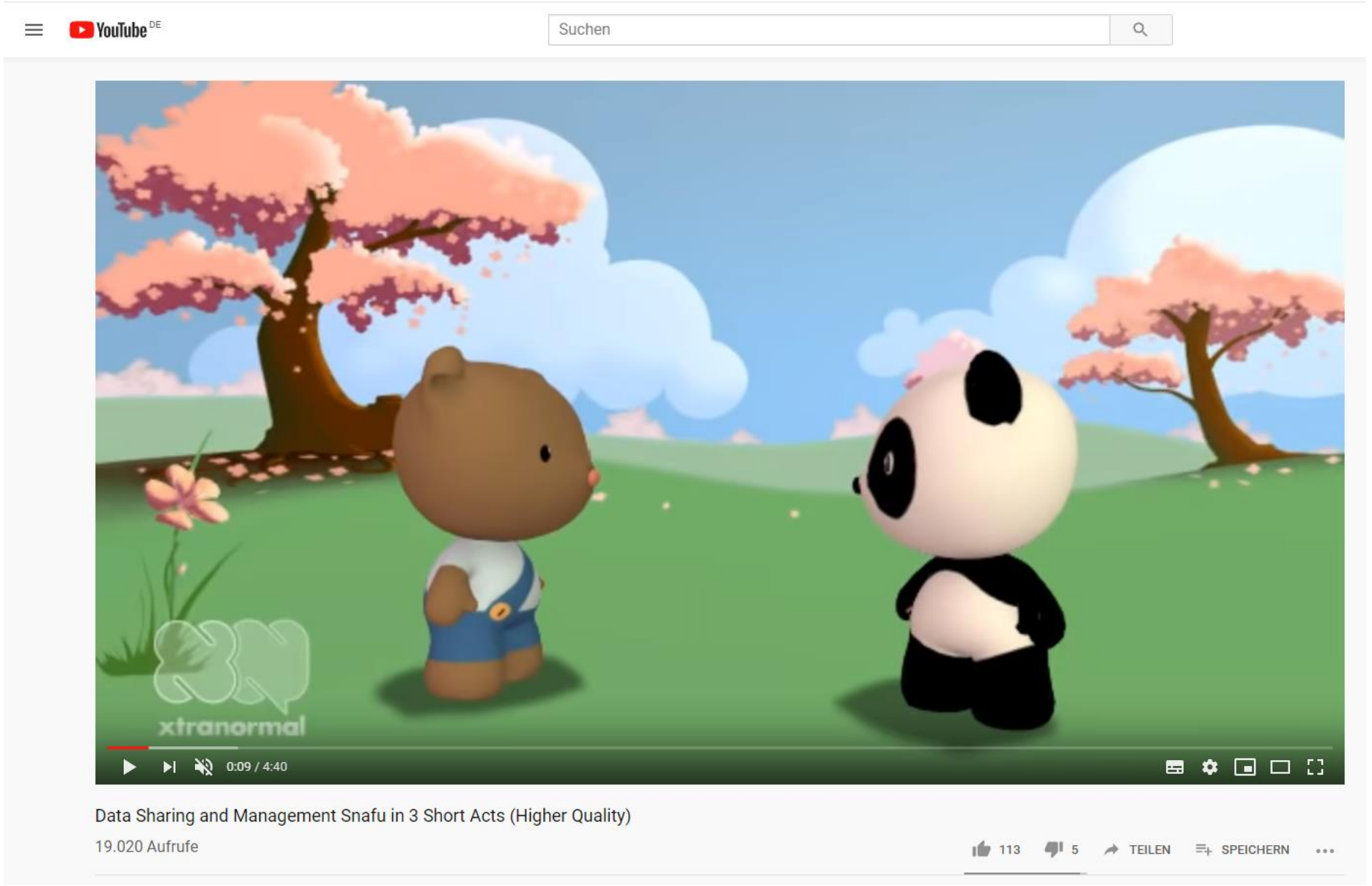
Roman Gerlach

Research Data Management Helpdesk
Michael-Stifel-Center Jena

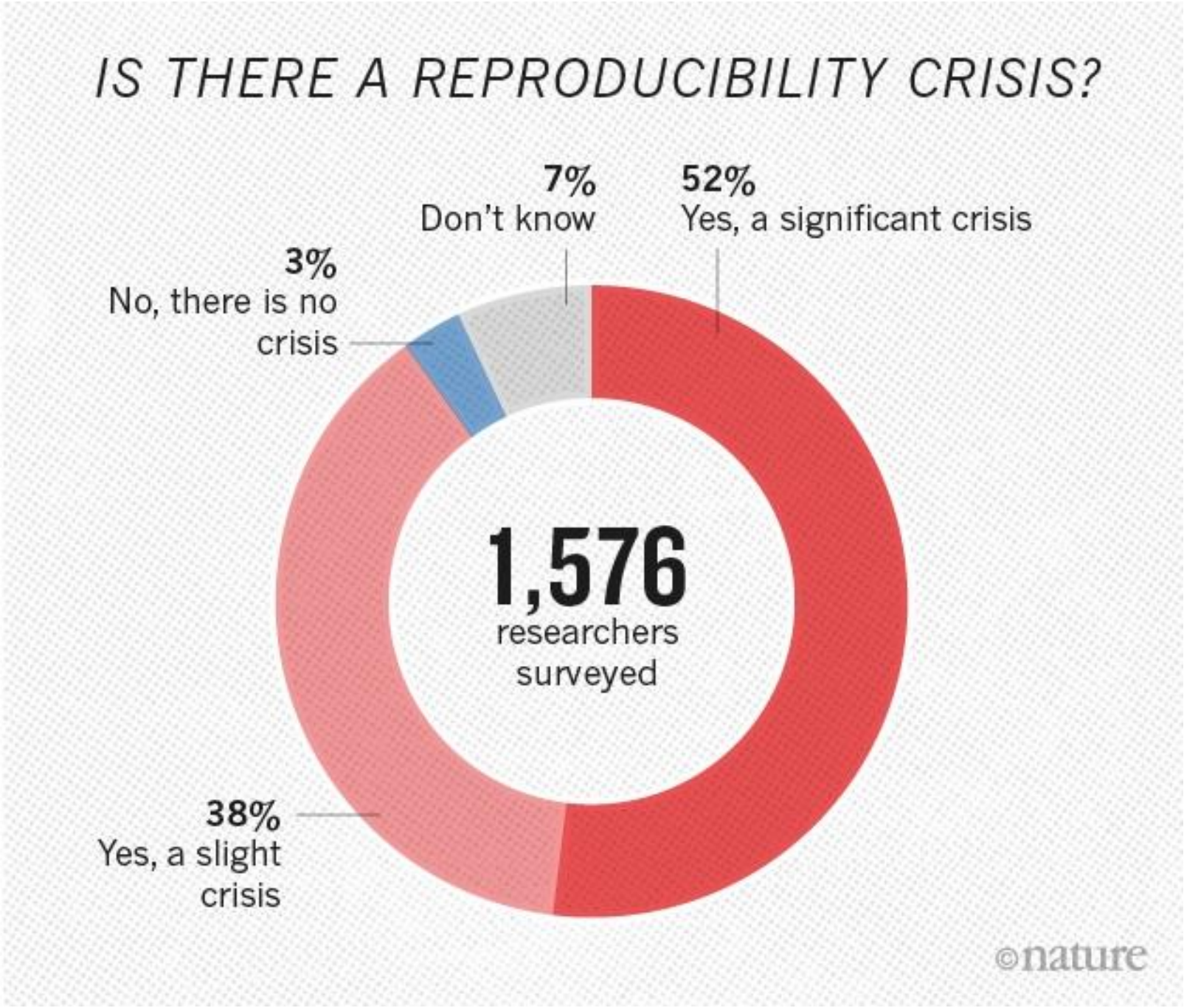
FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

Outline

1. Introduction / problem statement
2. How to do professional research data managment?
3. Documenting your data/work
4. Quality control and assurance
5. Data preservation and publication
6. Data management planing



https://www.youtube.com/watch?v=66oNv_DJuPc



Baker, M., Nature 533, 452–454 (26 May 2016) doi:10.1038/533452a

Question

What would you do to ensure others are able to reproduce your research findings?

The three pillars of reproducibility

- Documentation
- Quality Control & Assurance
- Preservation & Publication

Data Policies

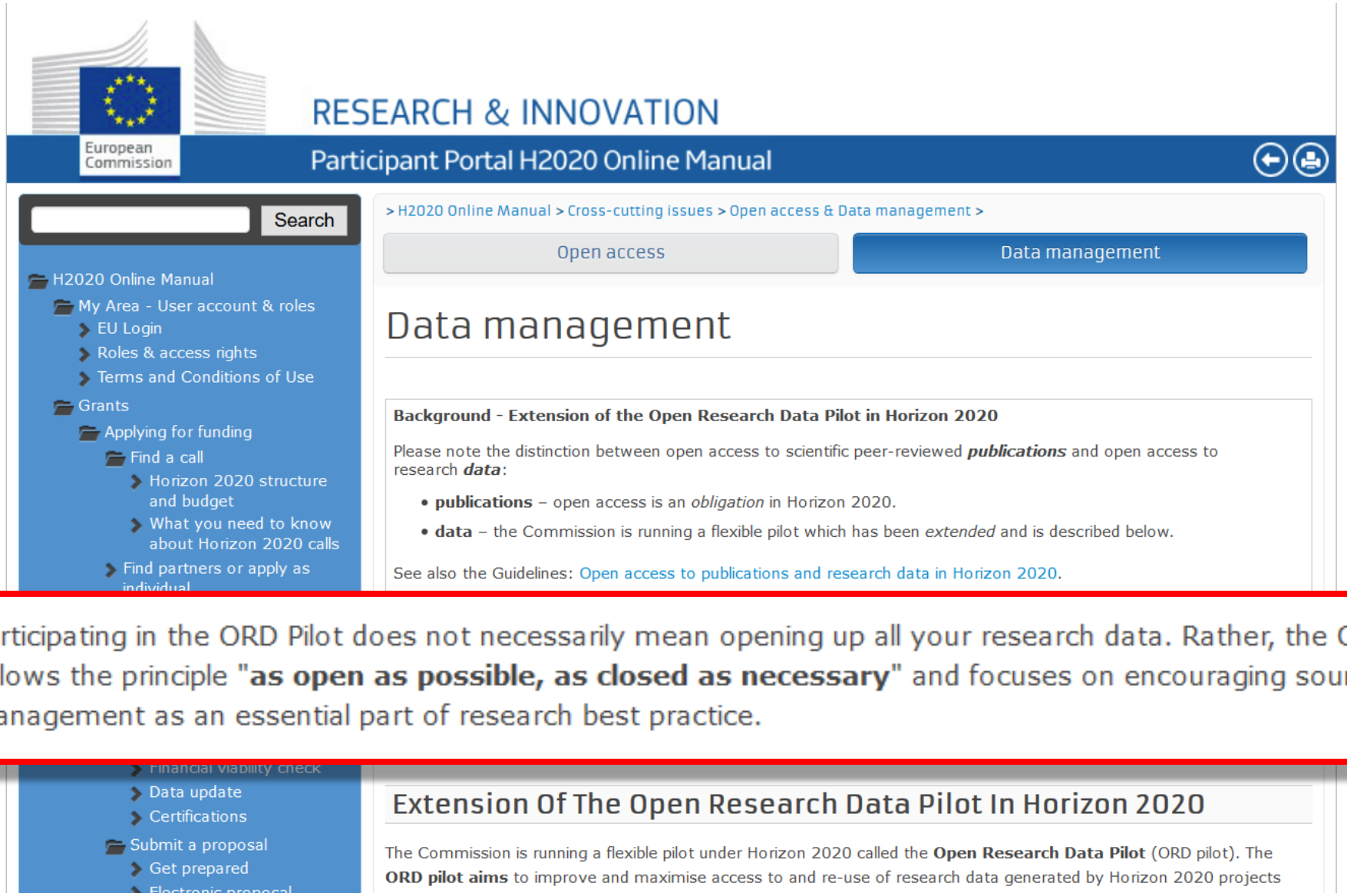
By international organisations

By funding agencies

By institutions

By journals

Example: EU



http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm

Example: DFG

Guidelines on the Handling Research Data (2015):

2. Accessibility
Assuming that the publication of research data from a DFG-funded project does not conflict with the rights of third parties (in particular data protection or copyright), research data should be made available as soon as possible. Data should be made accessible at a stage of processing that allows it to be usefully reused by third parties (raw data or structured data). To make sure this is the case, it must be ensured that access to the data is still guaranteed when, through publication, the rights of use relating to research data are transferred to a third party, usually a publishing house.

3. Long-term archiving
In accordance with the rules of good scientific practice, research data should be archived in the researcher's own institution or an appropriate nationwide infrastructure for at least 10 years.

More rules in Guidelines for proposals.

FAIR Principles

Findable

Accessible

Interoperable

Re-usable

https://www.force11.org/group/fairgroup/fairprinciples

FAIR Principles

- '**Findable**', i.e. discoverable with metadata, identifiable and locatable by means of a standard identification mechanism;
- '**Accessible**', i.e. always available and obtainable;
- '**Interoperable**', i.e. both syntactically parseable and semantically understandable, allowing data exchange and reuse between researchers, institutions, organisations or countries; and
- '**Reusable**', i.e. sufficiently described and shared with the least restrictive licences, allowing the widest reuse possible and the least cumbersome integration with other data sources.

https://www.slideshare.net/OpenAIRE_eu/horizon-2020-open-research-data-pilot-jeanclaude-burgelman-dg-rtd-european-commission-8th-openaire-workshop

Publishers Perspective

„[...] authors are required to make materials, data and associated protocols promptly available to readers without undue qualifications.“

“PLOS journals require authors to make all data underlying the findings described in their manuscript fully available without restriction, with rare exception.“

NPG (2013). Availability of data and materials. Retrieved from http://www.nature.com/authors/policies/availability.html
PLOS (2014). PLOS Editorial and Publishing Policies. Retrieved from http://www.plosone.org/static/policies.action

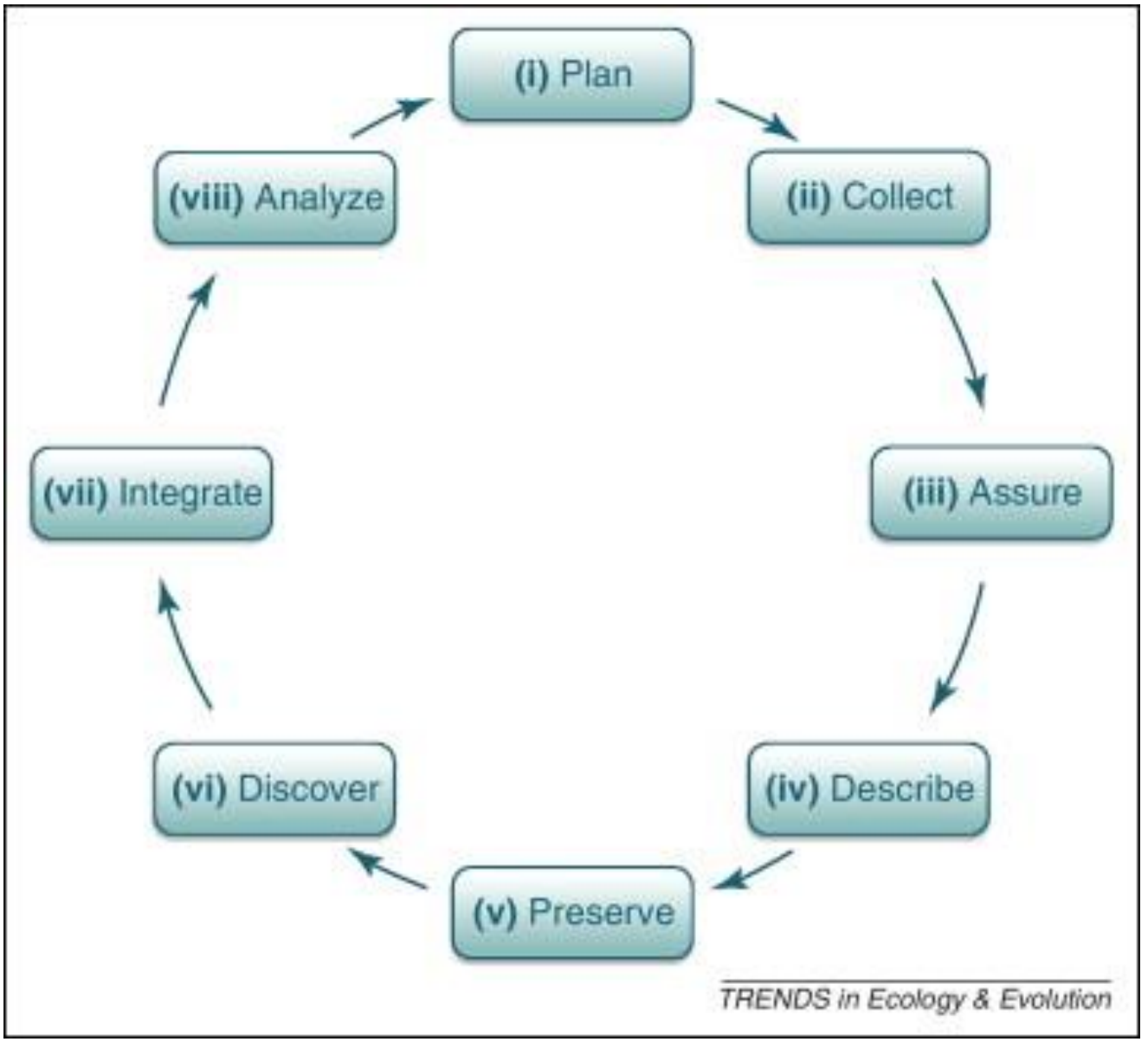
HELMHOLTZ GEMEINSCHAFT
Open Science

Folie: P. Vierkant (2015), Keynote: Freier Zugang zu Forschungsdaten

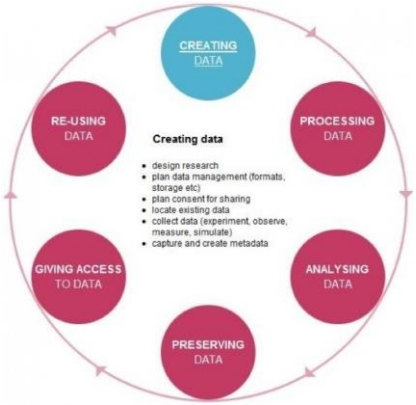
How to do professional research data managment?

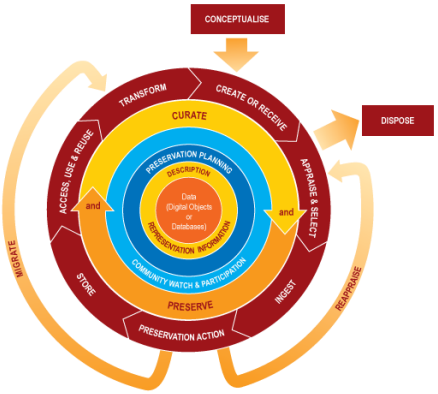
Recap: How does science work?

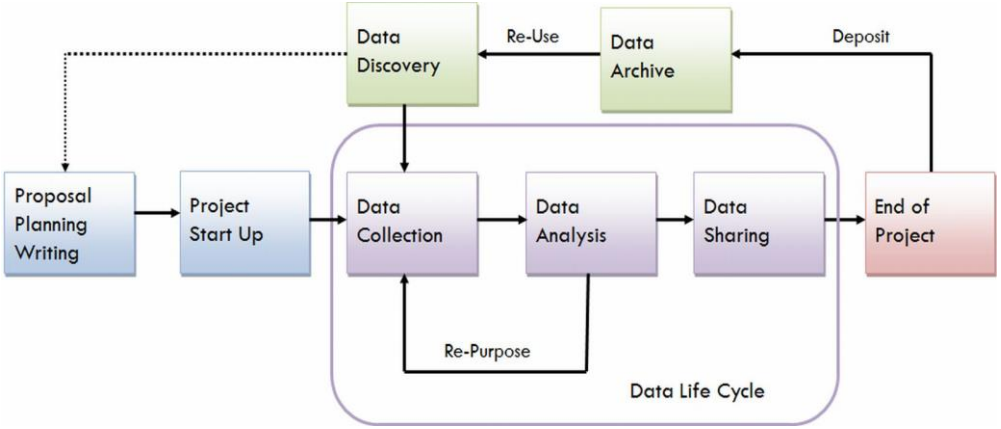
Data Life Cycle concept

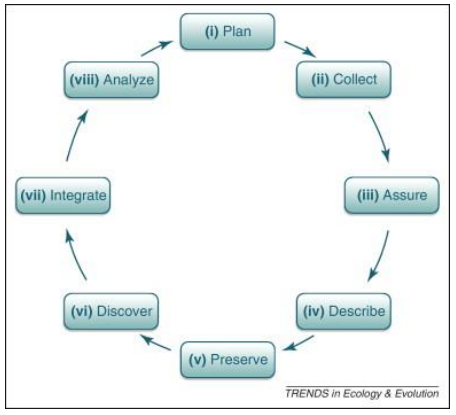


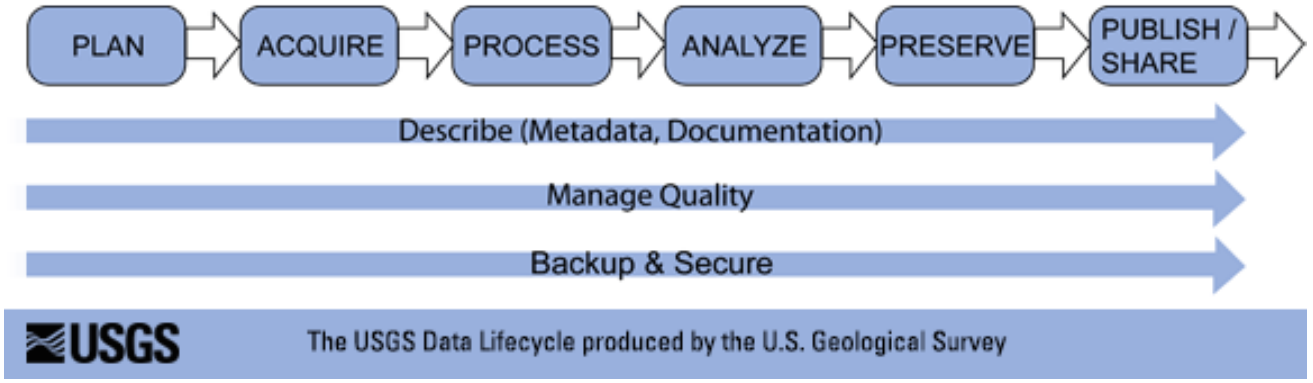
There are many lifecycles…











Documenting your work/data

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?

slide by DataONE

Metadata collection

Field/Lab book

human readable

Metadata / Lineage

```
<?xml version="1.0"?>
<dwr:DarwinRecordSet
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://rs.tdwg.org/dwc/dwcrecord/  http://rs.tdwg.org/dwc/xsd/tdwg_dwc_classes.xsd"
    xmlns:dcterms="http://purl.org/dc/terms/"
    xmlns:dwc="http://rs.tdwg.org/dwc/terms/"
    xmlns:dwr="http://rs.tdwg.org/dwc/dwcrecord/">
    <dcterms:Location>
        <dwc:locationID>http://guid.mvz.org/sites/arg/127</dwc:locationID>
        <dwc:country>Argentina</dwc:country>
        <dwc:countryCode>AR</dwc:countryCode>
        <dwc:stateProvince>Neuquén</dwc:stateProvince>
        <dwc:locality>25 km al NNE de Bariloche por Ruta 40 (=237)</dwc:locality>
    </dcterms:Location>
    <dwc:Occurrence>
        <dcterms:type>PhysicalObject</dcterms:type>
        <dcterms:modified>2009-02-12T12:43:31</dcterms:modified>
        <dcterms:rightsHolder>Museum of Vertebrate Zoology</dcterms:rightsHolder>
        <dcterms:rights>Creative Commons License</dcterms:rights>
```

human AND machine readable
+ searchable
+ interconnected (linked data)
+ standardized

How to create quality metadata?

Tips 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)

Tips for Writing Quality Metadata

A Clear Choice: Which title is better?

Rivers

OR

Greater Yellowstone Rivers from 1:126,700 U.S. Forest Service Visitor Maps (1961-1983)

Greater Yellowstone (where) Rivers (what) from 1:126,700 (scale) U.S. Forest Service (who) Visitor Maps (1961-1983) (when)

Tips for Writing Quality Metadata

Use standards whenever possible:

ISO 639 Code for the representation of the **names of languages**.
e.g. eng – English; ger/deu – German; fre/fra - French
http://www.loc.gov/standards/iso639-2/langhome.html

ISO 3166 Codes for the representation of **names of countries**.
e.g. DE – Germany; IR – Iran; CN – China; HR - Croatia
http://www.oasis-open.org/cover/country3166.html

ISO 8601 Codes for the representation of **date and time**.
e.g. 2014-06-19T13:15:30Z; 2014-06-19T13:16:30+01:00
http://en.wikipedia.org/wiki/ISO_8601

Tips for Writing Quality Metadata

- Remember: a computer will read your metadata
- Do not use symbols that could be misinterpreted: Examples:

  ! @ # % { } | / \ < > ~
- Don’t use tabs, indents, or line feeds/carriage returns
- When copying and pasting from other sources, use a text editor (e.g., Notepad) to eliminate hidden characters

The 20-Year Rule

- The metadata accompanying a data set should be written for a user 20 years into the future--what does that investigator need to know to use the data?
- Prepare the data and documentation for a user who is unfamiliar with your project, methods, and observations

NRC (1991)

The three pillars of reproducibility

- Documentation 
- Quality Control & Assurance
- Preservation & Publication

Quality Control & Assurance

Data quality

What does quality assurance mean?

- to make sure others can understand your data (i.e. your conclusions)
- to enable others to assess **fitness for use**

Reference: A.D. Chapman, 2005, Principles of Data Quality

Data quality

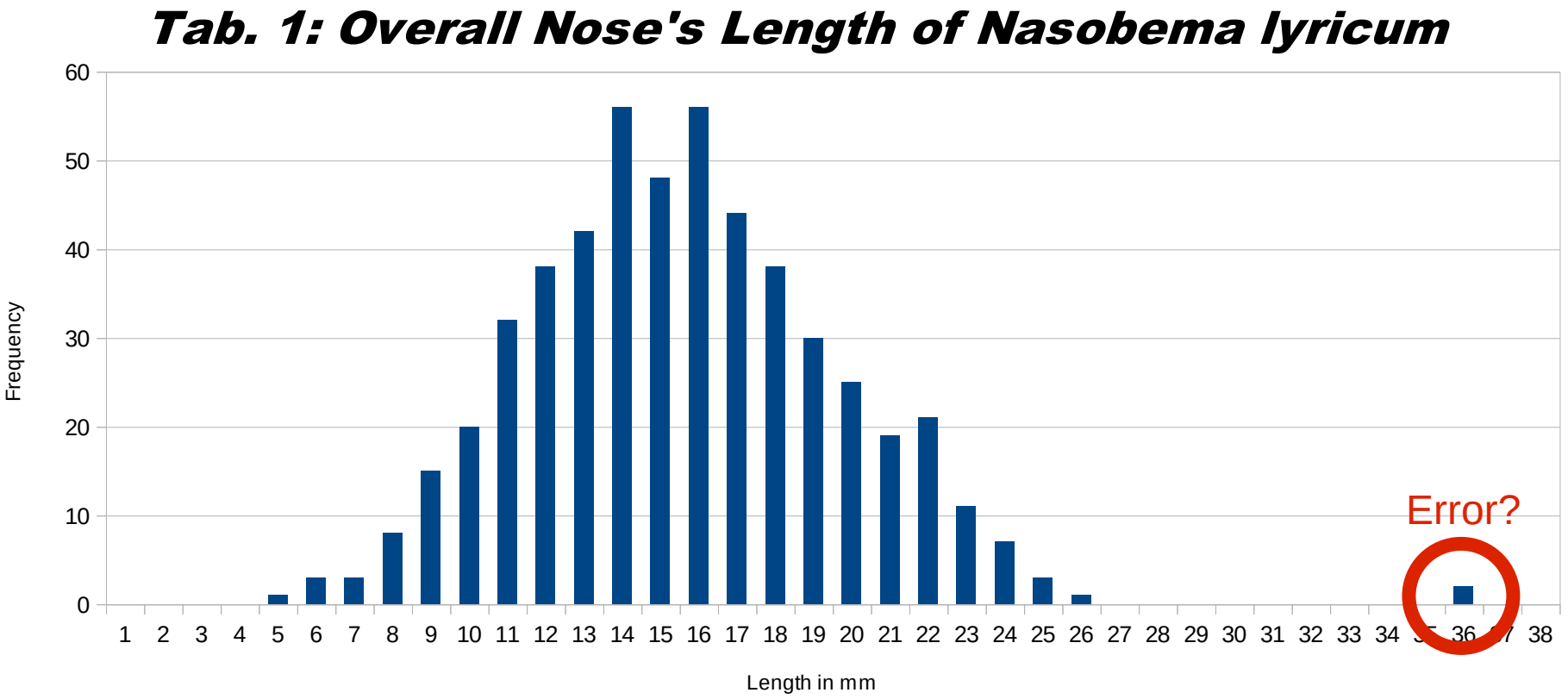
How to assure data quality?

- Verification and review
- Ensure consistency, accuracy, completeness, reliability, …
- providing context information = documentation, provenance

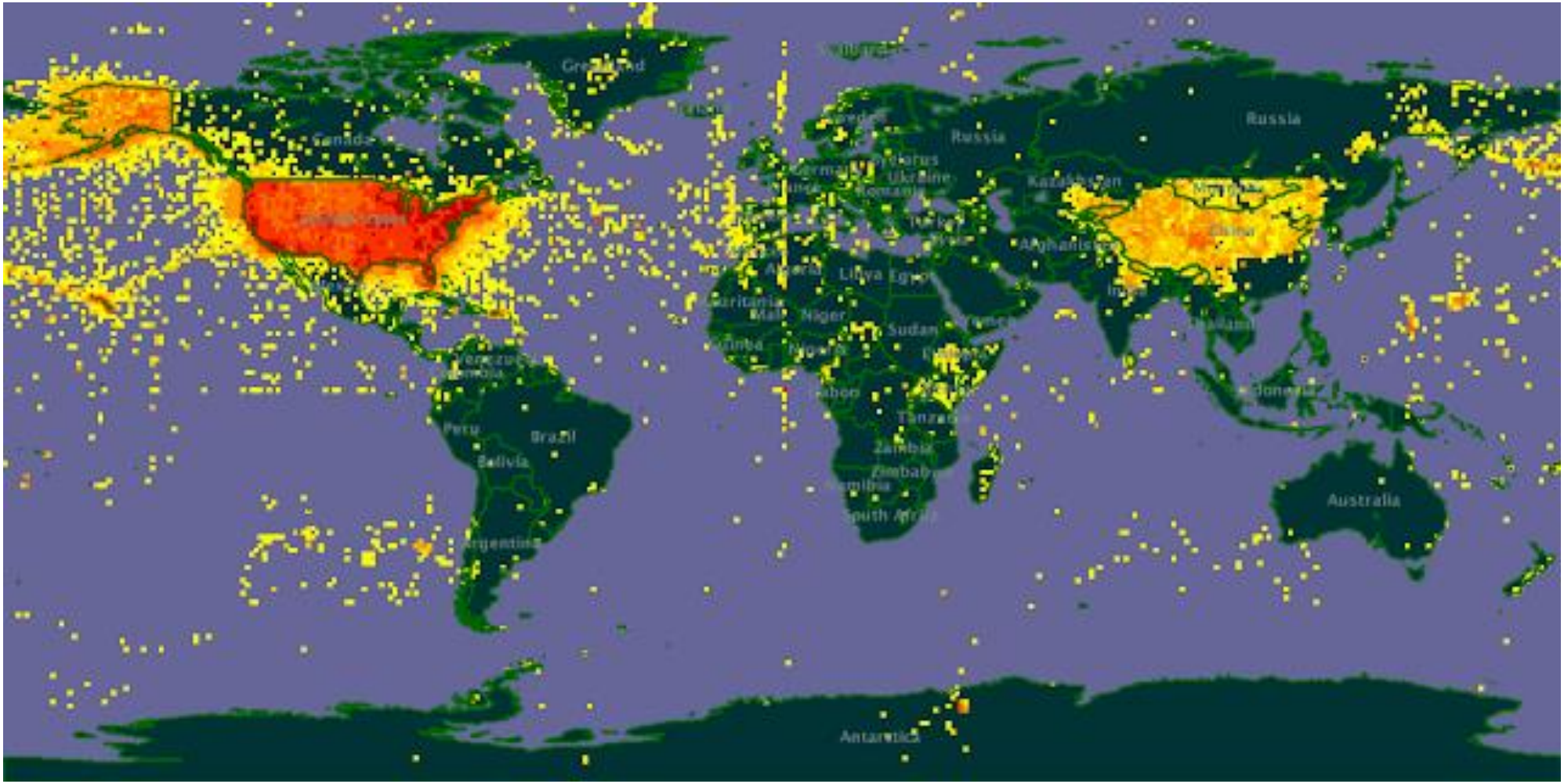
Reference: A.D. Chapman, 2005,
Principles of Data Quality

Error detection

Visualize data values

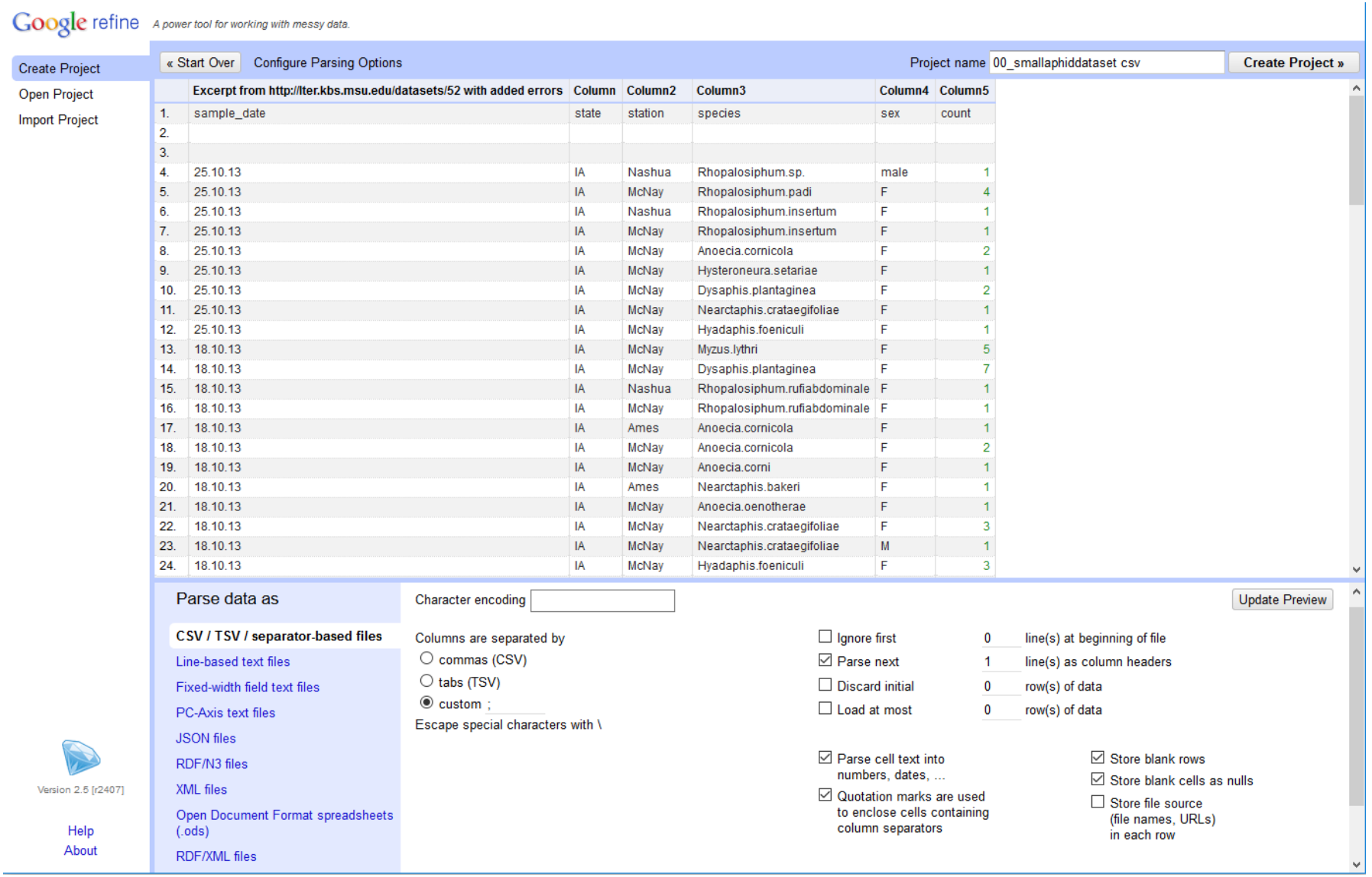


Visualization



Place geographic data on a map to ensure that geographic coordinates are correct.

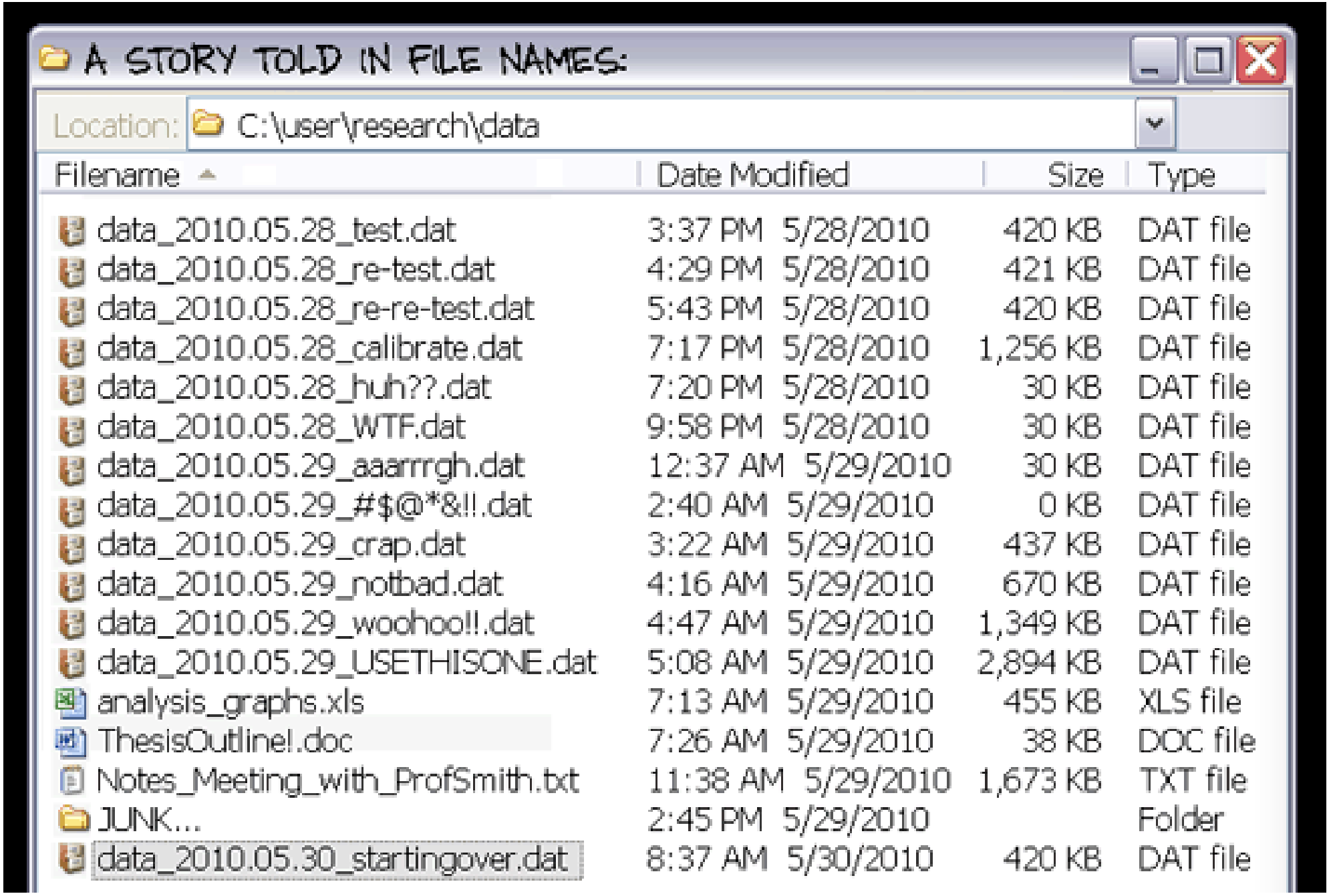
Tool: OpenRefine



FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

http://openrefine.org

File Naming



http://phdcomics.com/comics.php?f=1323

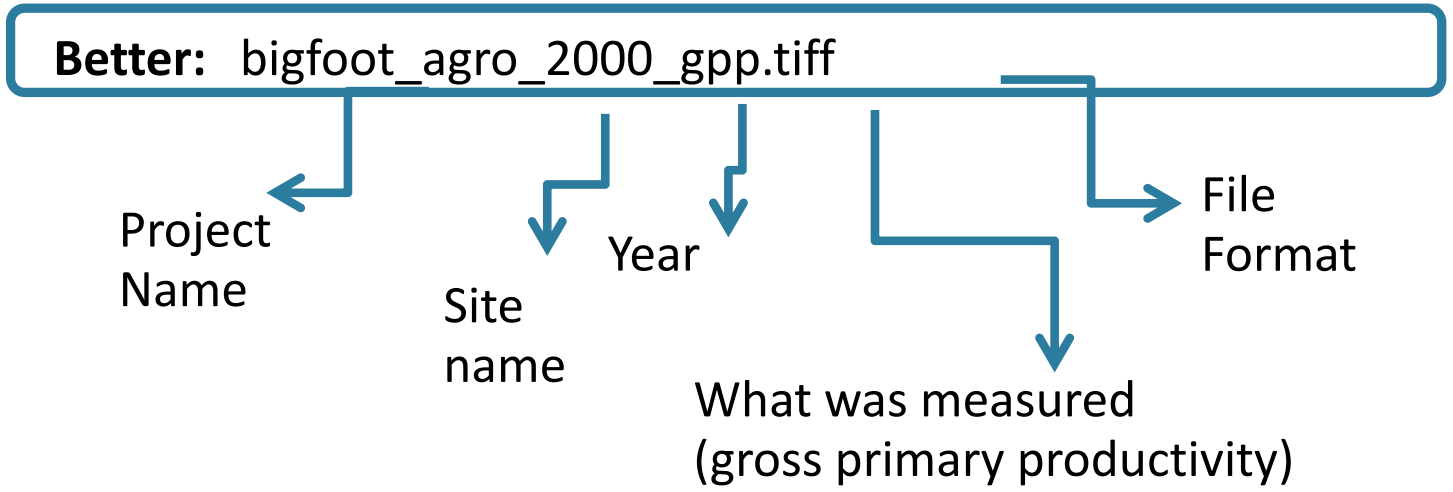
File Naming

- Use descriptive file names
  - Unique
  - Reflect contents
  - ASCII characters only
  - Avoid spaces

Bad: Mydata.xls

2001_data.csv
best version.txt

Better: bigfoot_agro_2000_gpp.tiff



File Naming

- Be **consistent** throughout your project directories/files
- Keep it **short!**
- A good format for **date designations** is YYYYMMDD_...
- use "001, 002, ...010, 011, etc." instead of "1, 2, ...10, 11 ... 100, 101, etc.“

Folder Structure

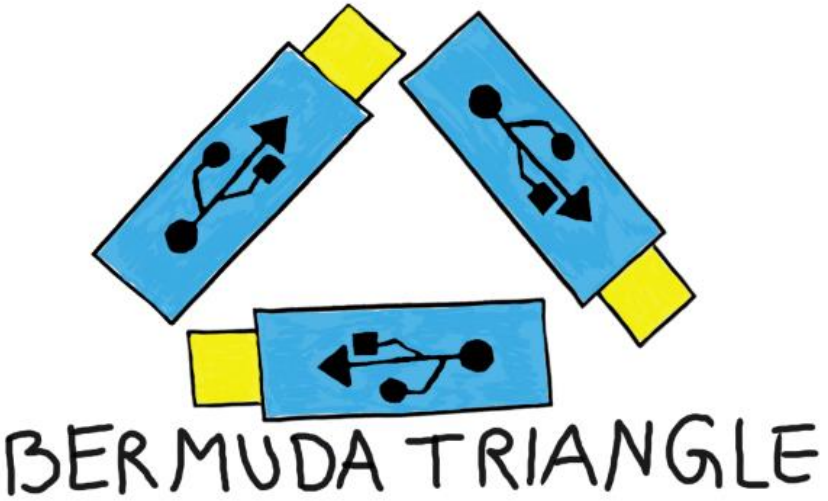
The **primary goals** of the folder structure are to

- Provide **predictable locations** for data files and other types of resources (e.g. code)
- Provide distinction between ”**original**”,”**finalized”** datasets and datasets that are still in production
- Provide a **space for archiving** outdated datasets
- for each directory include a **readme.txt** file that explains your naming format along with any abbreviations or codes you have used.

The three pillars of reproducibility

- Documentation 
- Quality Control & Assurance 
- Preservation & Publication

Data Preservation



Data Preservation

Or: How to make sure that data remains available after your project ends, after you leave the university, after your boss retires,…

Data preservation

Or: How to make sure that data remains available after your project ends, after you leave the university, after your boss retires,…

PALEOFUTURE

SPACE

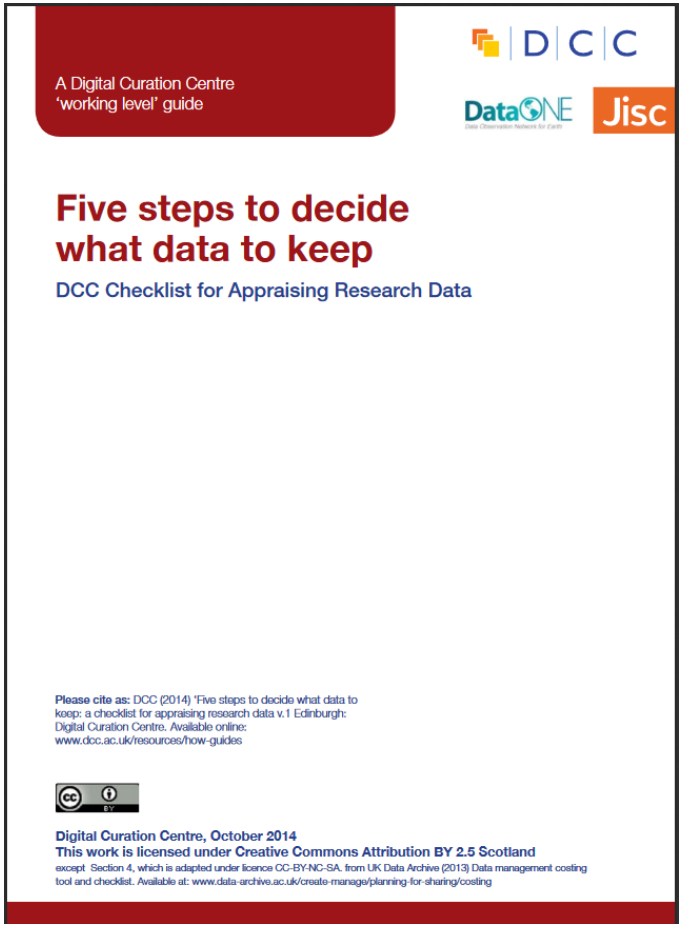
McMoon's: The Former McDonald's Where NASA Digitizes Old Moon Photos

Matt Novak
8/19/13 12:59pm • Filed to: SPACE

9 2

http://www.moonviews.com/

Five steps to decide what data to keep



What data and for how long?

Step 1. Identify **purposes** that the data could fulfill

Step 2. Identify data that **must** be kept

Step 3. Identify data that **should** be kept

Step 4. Weigh up the **costs**

Step 5. **Complete** the data appraisal

http://www.dcc.ac.uk/resources/how-guides/five-steps-decide-what-data-keep

Identify data that must be kept

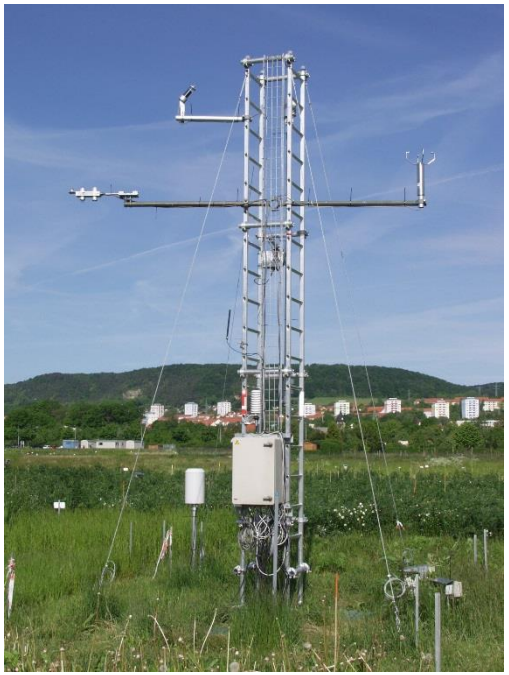
- Are there Research Data Policy reasons to keep it?
- Do regulations require the data to be available?
- Are there other legal or contractual reasons?
- Does it contain personal data relevant to the reuse purpose?

Identify data that should be kept

- Is it good enough?
- Is there likely to be a demand?
- How difficult is it to replicate?
- Do any barriers to further use exist?
- Is it the only copy?

Data appraisal

1. data that cannot be recreated, because it contains unique observations/measurements in space and time



Data appraisal

2. data that can only be reproduced with enormous effort (e.g. super computer calculations)

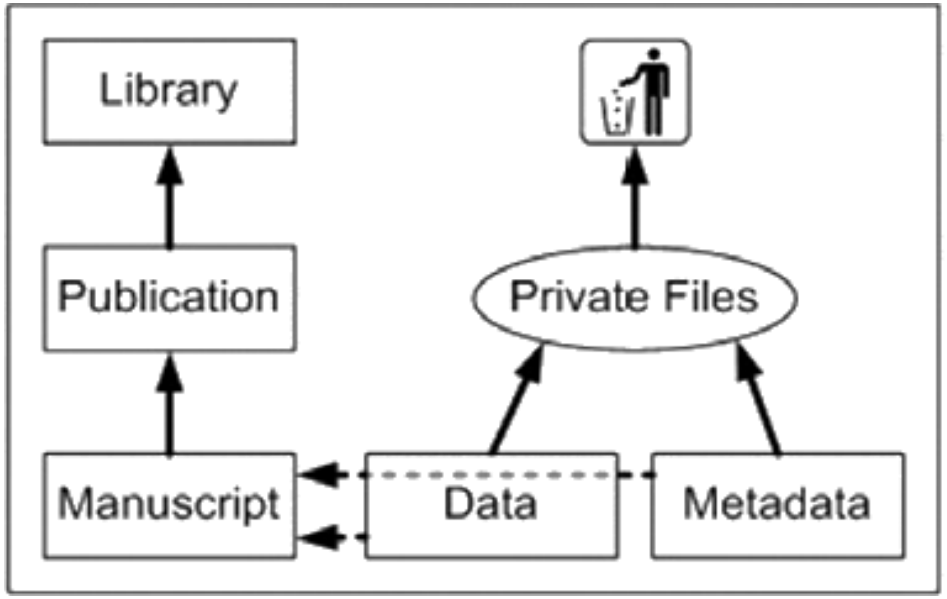
Long-term archiving

Challenges:

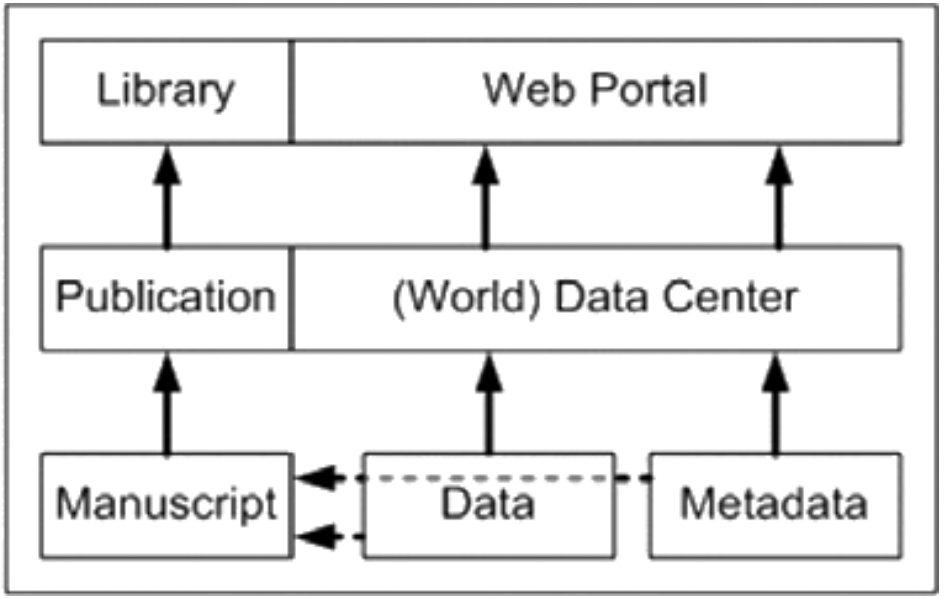
- File formats
- Migration (Hardware, Software, Format, …)
- Data retirement/disposal
- Continuous funding

Data Publication

Data Publication today!?



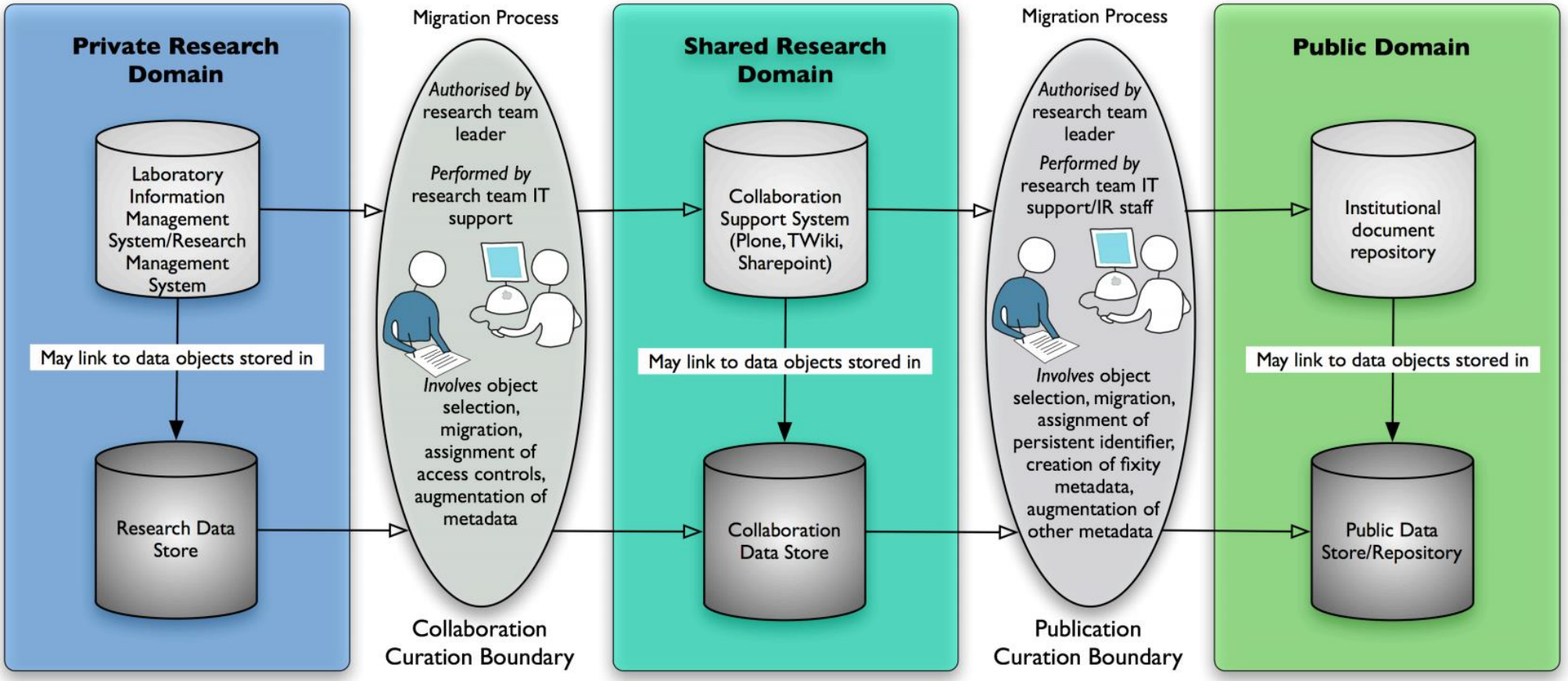
Common way



Prefered way

Rümpel 2011:26

Data Domains



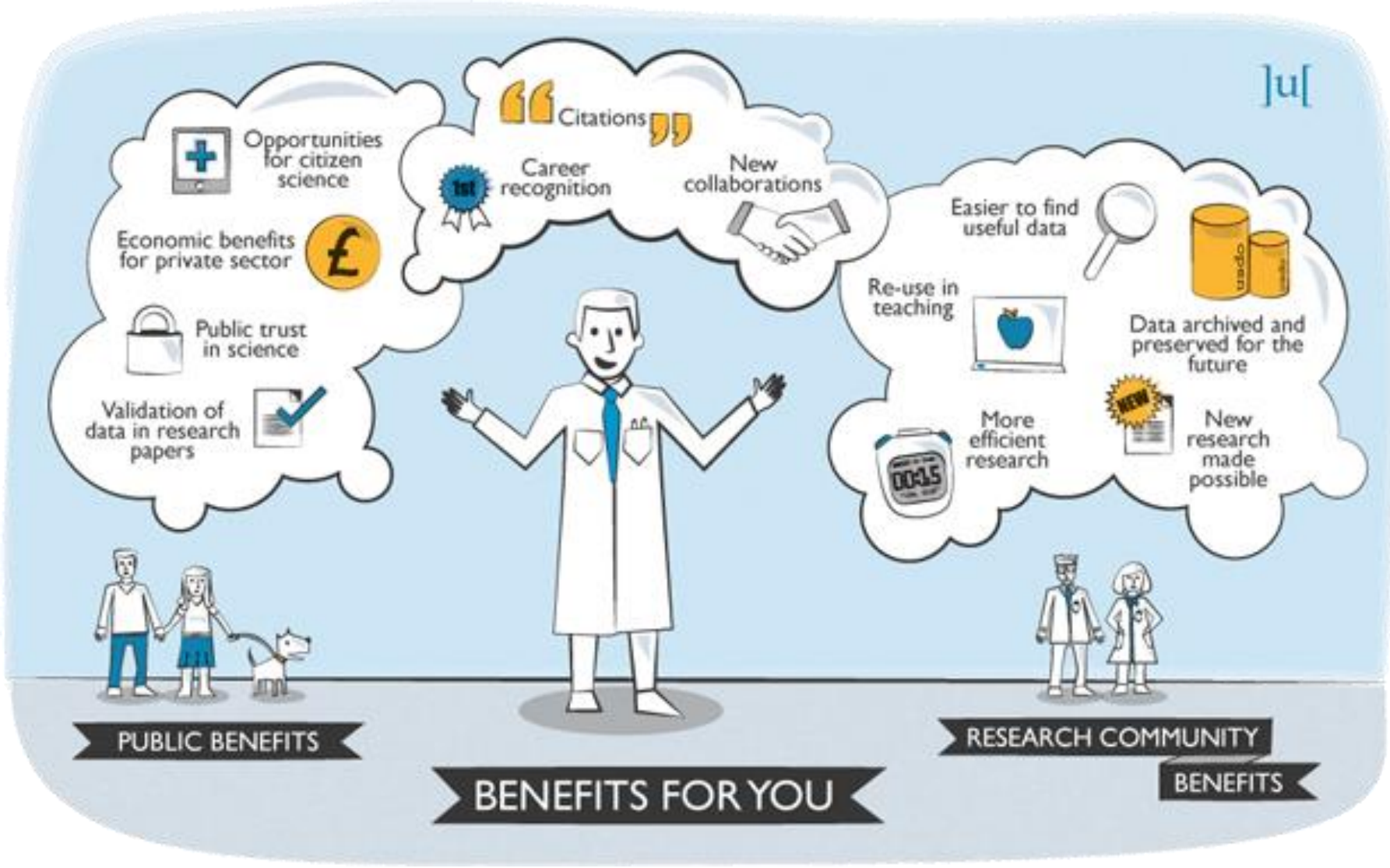
This domain involves the core research team as they undertake the research, usually within a single institution. Access is often tightly controlled as hypotheses and analyses are developed.

This domain involves researchers outside the core team as they collaborate with colleagues, often across institutions. Access is more open, but not everything is shared.

This domain involves the public sphere (publication in the sense of making public). Access will usually be open to all.

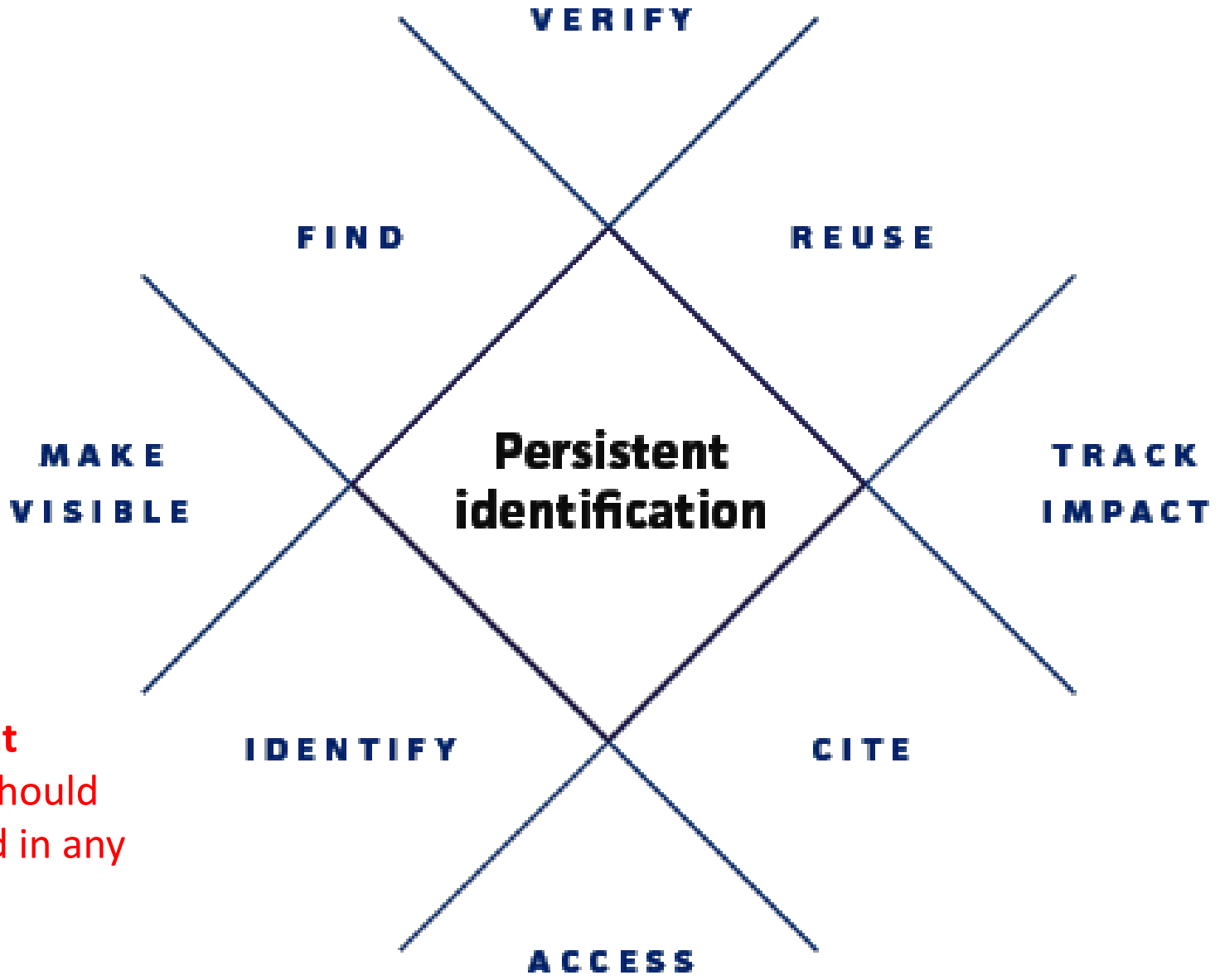
Version 1.4, http://andrew.treloar.net/, 07Dec07

Benefits of data publication



How to find something after years in the internet?

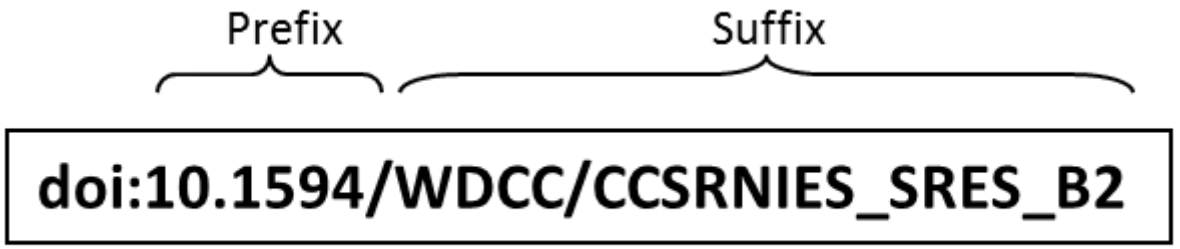
Persistent Identifier



A **persistent identifier** should be included in any citation

PID Example: DOI

- **DOI** (Digital Object Identifier)
  - Globally unique, alphanumeric string assigned by a registration agency



  - DOI name should **not** contain any changeable attributes of the object (e.g. physical location, ownership); attributes are encoded in the metadata, not in the DOI name
  - A DOI should always resolve to a **landing page**

PID Example:

ORCID
Connecting Research and Researchers

Researcher identifier: **ORCID** (Open Researcher & Contributor ID)

- Central registry of unique identifiers for individual researchers to address author name ambiguity
- Transparent linking mechanism between ORCID and other author ID schemes
- Example: orcid.org/0000-0001-5104-4247

Where can I publish my data?

- Journal paper supplement
- Data Journal
- Data Repository

Journal

Journal = Paper + PID + [supplement]

Data Journal

Data Journal = Paper + Datasets + PID

An 18-yr long (1993–2011) snow and meteorological dataset from a mid-altitude mountain site (Col de Porte, France, 1325 m alt.) for driving and evaluating snowpack models

S. Morin, Y. Lejeune, B. Lesaffre, J.-M. Panel, D. Poncet, P. David, and M. Sudul

Page(s) 13-21

Abstract | Final Revised Paper (PDF, 2559 KB) | Supplement (14997 KB) | Discussion Paper (ESSDD)

WILEY ONLINE LIBRARY
RMetS Royal Meteorological Society
Geoscience Data Journal

CODATA DATA SCIENCE Journal

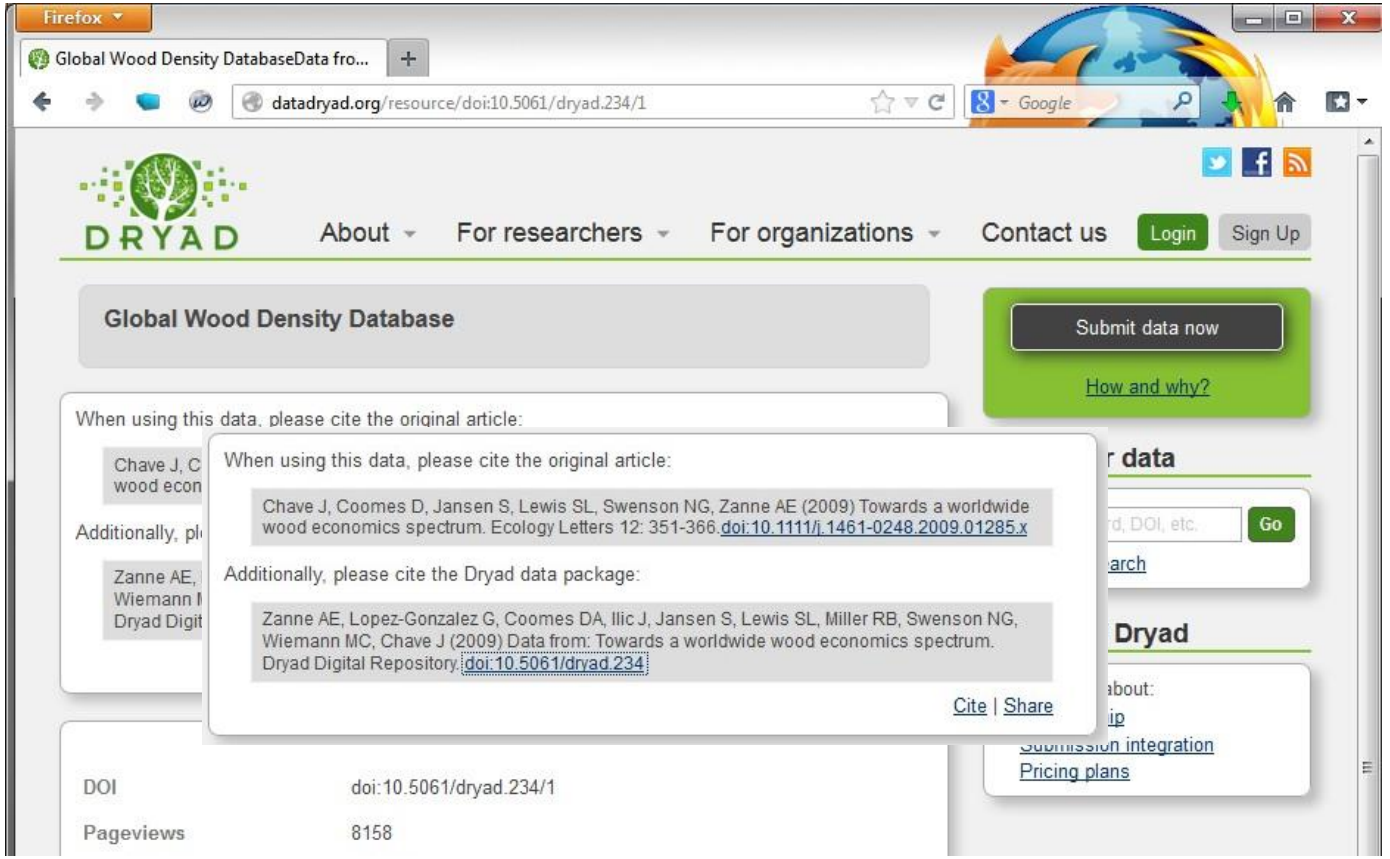
A peer-reviewed open-acces
Biodiversity Data Journal
Making your data count!

npg nature publishing group
SCIENTIFIC DATA

List of Data Journals:
http://www.researchdata.uni-jena.de/Datenpublikation.html

Data repository

Data repository = Datasets + Metadata + PID

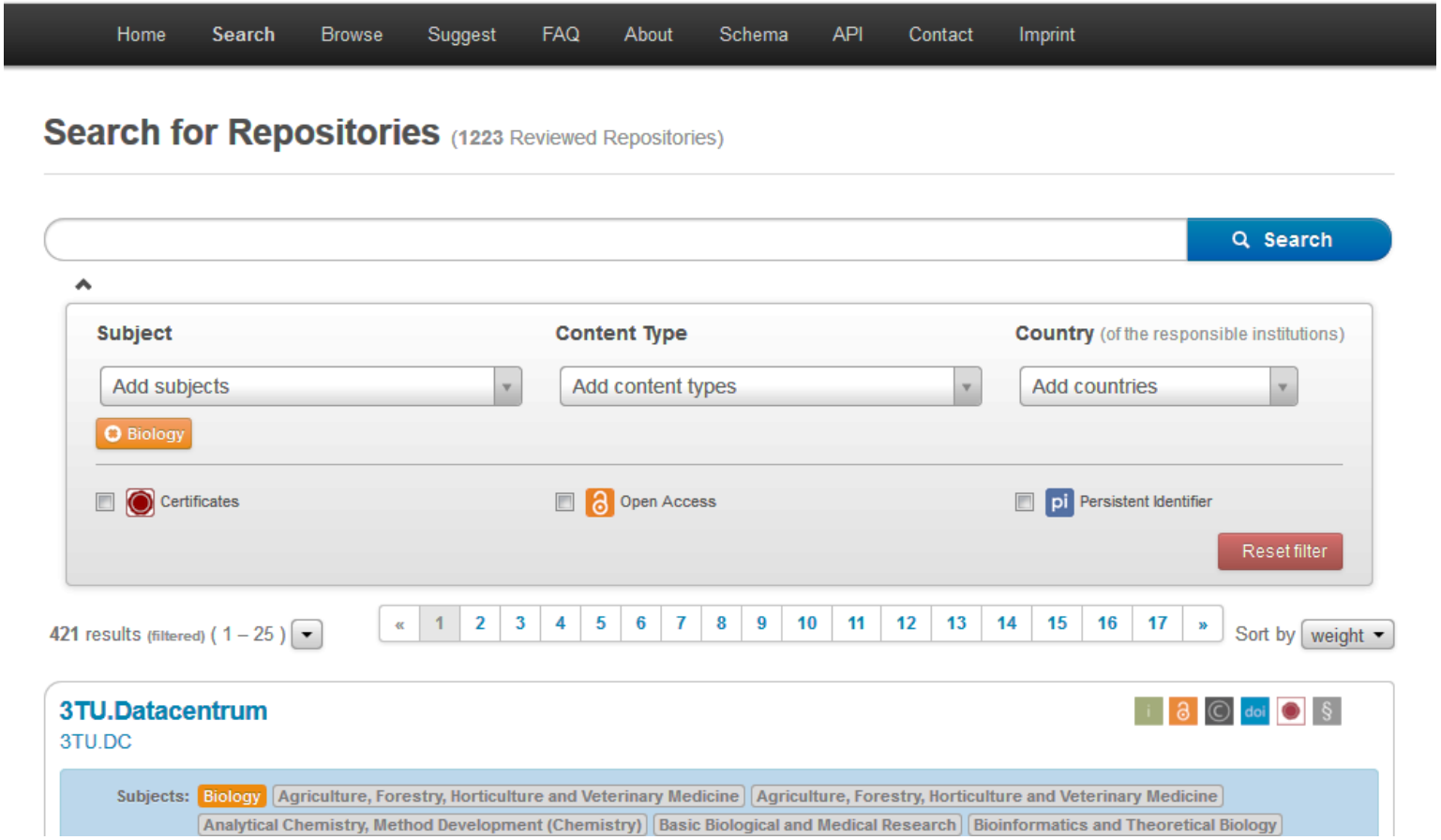


Types of Repositories

- Domain specific repositories
- Generic/institutional repositories

How to find a data repository?

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

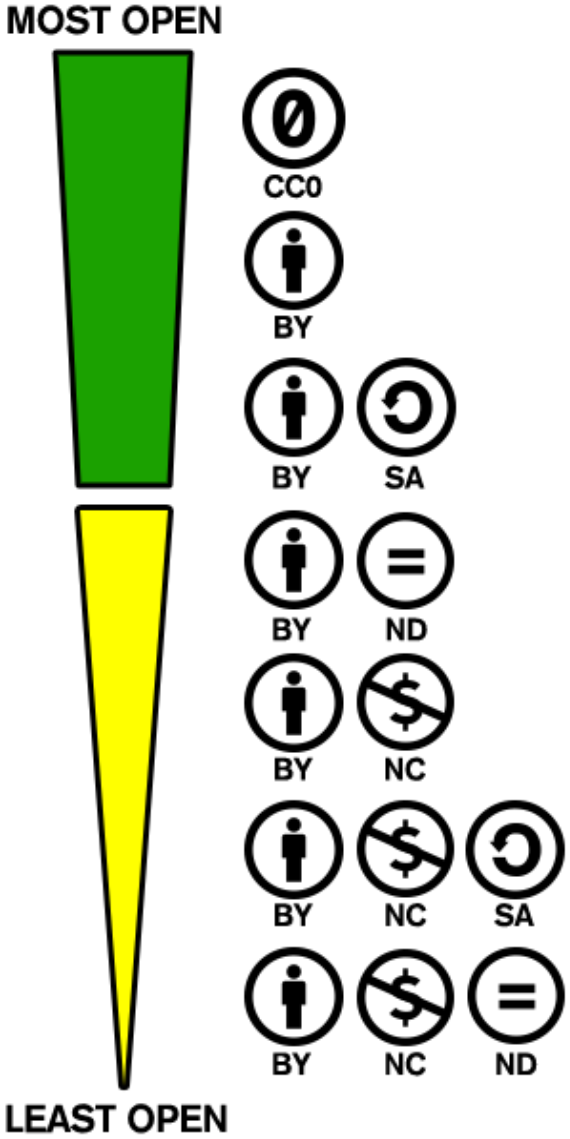


How to select a repository

Criteria:

- Long-term availability (Certification?)
- Provides PIDs (e.g. DOI, URN, ARK)
- License (e.g. Open Access, Restricted)
- 1st domain, 2nd generic
- Data review process?

Licences: e.g. Creative Commons



Only version 4.0 of

- CC0
- CC-BY
- CC-SA

conform to the Open Definition (i.e. Open Access)

Summary: Knowledge is open if anyone is free to access, use, modify, and share it — subject, at most, to measures that preserve provenance and openness.

http://opendefinition.org/od/2.1/en/

Data Management Planning

1. How will the data be managed during the project?

Data Storage

- Is all data **digital**?
- What data **volume** will be produced? More than 500GB?
- Do you have a **naming convention**?
- What **software** and **file formats** will you use?

Photo by Drew Hays on Unsplash-206414

Data description and documentation

- What standards will be used to record the data?
- What additional information is required?
  - Collection methods
  - Context information (metadata)
- Where will the metadata be located

Photo by Stefan Stefancik on Unsplash-257625

Data Access & Security

- Version control (e.g. SVN, Git)
- Back-ups (frequency, media)
- Security & protection from manipulation/theft
- Who will be responsible?

2. How will the data be managed after the project?

Plans for longterm archiving

- Where will the data be archived?
- For how long?
- What data should be archived?
- What data can be deleted?

Photo by Sanwal Deen on Unsplash-93466

Data Sharing & Publication

- How can I share my data? Isn't my Website enough?
- Where can I publish my data (e.g. data journal, research data repository)?
- Where can I get persistent identifiers (PID) for my data (e.g. DOI, URN)?
- What data objects get a PID (i.e. data granularity)?

Other Aspects of Data Management Planning

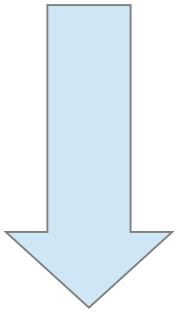
Legal & Ethical Aspects

- Ownership, copyright, intellectual property, licenses
- Are there ethical issues with the data (e.g. privacy, confidentiality, cultural sensitivity)
- Do I need written agreements from probants/patients?
- Do I need to anonymize the data?
- Do we need a data policy for our project?

Planning cost

- **Data collection** cost (e.g. data purchase, hardware, labour, usage of equipment)
- **Data infrastructure** cost (i.e. storage, computing power)
- **Labour** cost (e.g. data manager, system admin, Hiwi's)
- **Publication** cost (e.g. Open Access, archiving)

To ensure that no step of research data lifecycle is forgotten



Research Data Management Plan

(DMP)

What is a DMP?

- (formal) document that outlines how you will handle your data both during your project, and after the project is completed
- It's a living document

Data Management Plan for Post-Graduate Research Projects

Researcher:	
Project Title:	
Project Duration:	
Project Context:	
1. What Data will be Produced?	
2. How will the Data be Documented and Described?	
3. Has a 'File Structure/Naming Form' been completed?	
4. Deposition of E-Thesis:	
5. What are the plans for data sharing and access after submission of the thesis?	
6. What are the plans for long-term archiving of the digital data supporting the thesis?	
Signed:	**Version:**
Date Created:	**Date Amended:**

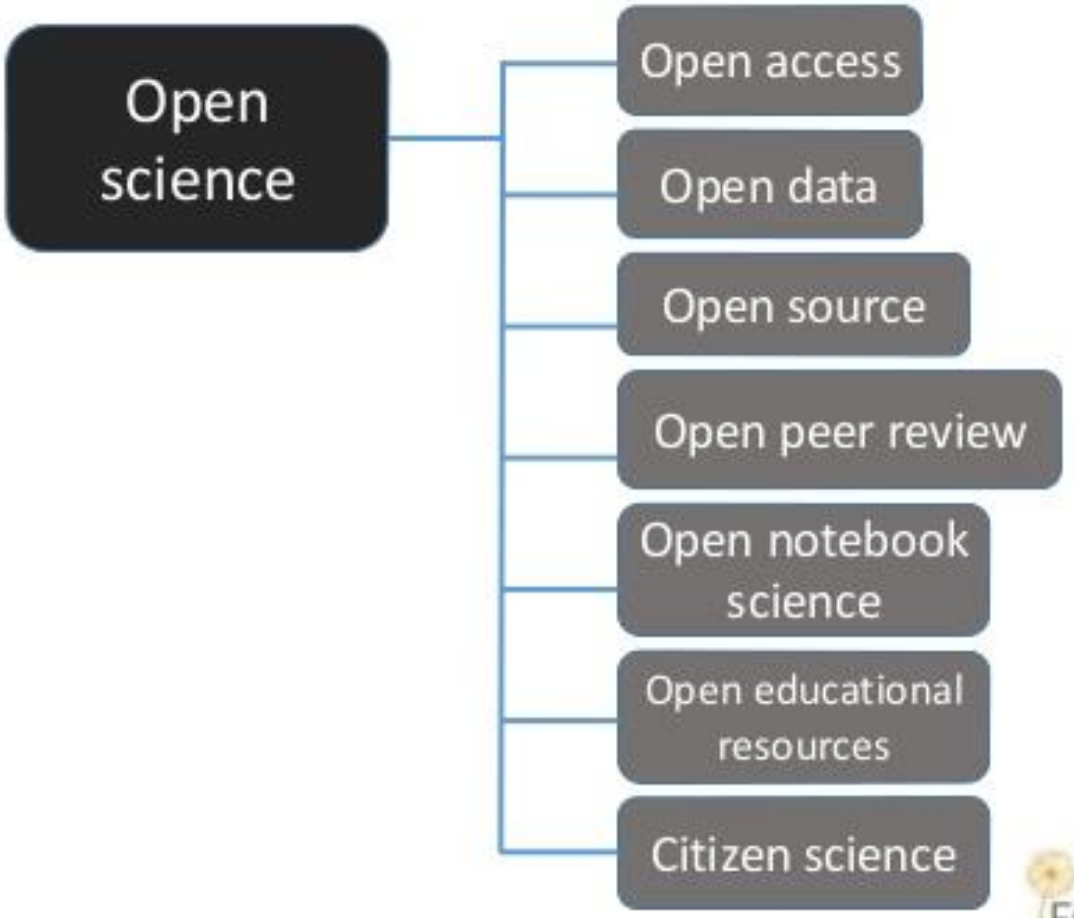
For your project

1. Make a plan! (incl. back-up & sharing)
2. Document your processing steps
3. Make sure your data comes with metadata
4. Implement consistent file naming (+ versioning)
5. Prepare your data/project so that others can understand and use it

Thanks for attending!

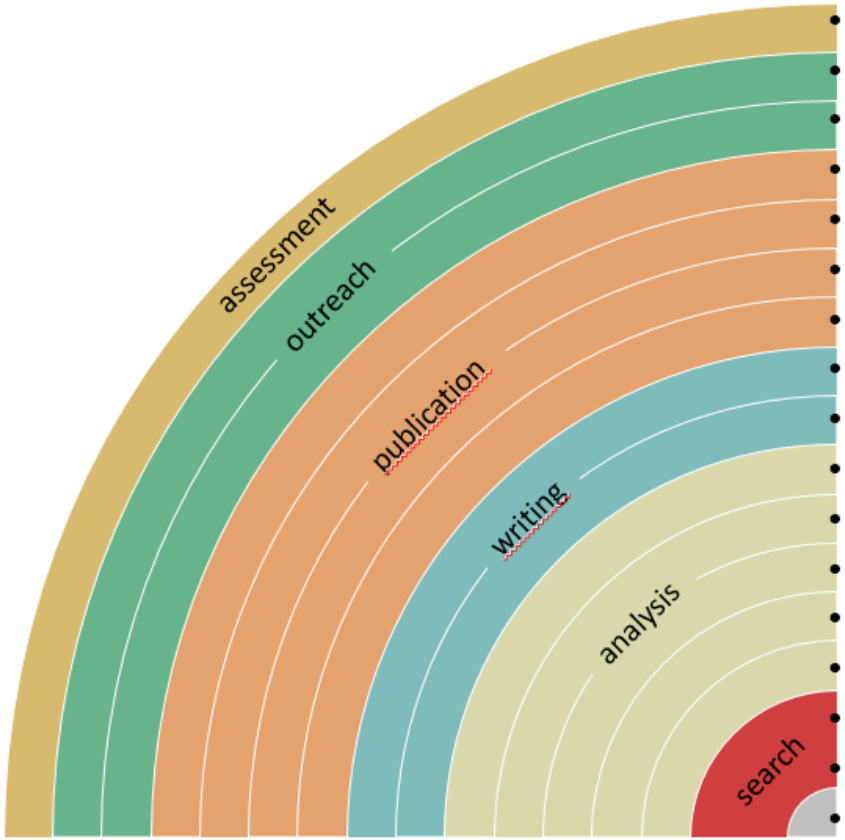
Backup slides

Open Science



Source: https://image.slidesharecdn.com/potsdam2015-lsu-160111213959/95/open-science-open-data-foster-training-potsdam-6-638.jpg

You can make your workflow more open by …



- adding alternative evaluation, e.g. with altmetrics
- communicating through social media, e.g. Twitter
- sharing posters & presentations, e.g. at FigShare
- using open licenses, e.g. CC0 or CC-BY
- publishing open access, 'green' or 'gold'
- using open peer review, e.g. at journals or PubPeer
- sharing preprints, e.g. at OSF, arXiv or bioRxiv
- using actionable formats, e.g. with Jupyter or CoCalc
- open XML-drafting, e.g. at Overleaf or Authorea
- sharing protocols & workfl., e.g. at Protocols.io
- sharing notebooks, e.g. at OpenNotebookScience
- sharing code, e.g. at GitHub with GNU/MIT license
- sharing data, e.g. at Dryad, Zenodo or Dataverse
- pre-registering, e.g. at OSF or AsPredicted
- commenting openly, e.g. with Hypothes.is
- using shared reference libraries, e.g. with Zotero
- sharing (grant) proposals, e.g. at RIO

DOI: 10.5281/zenodo.1147025