CSC

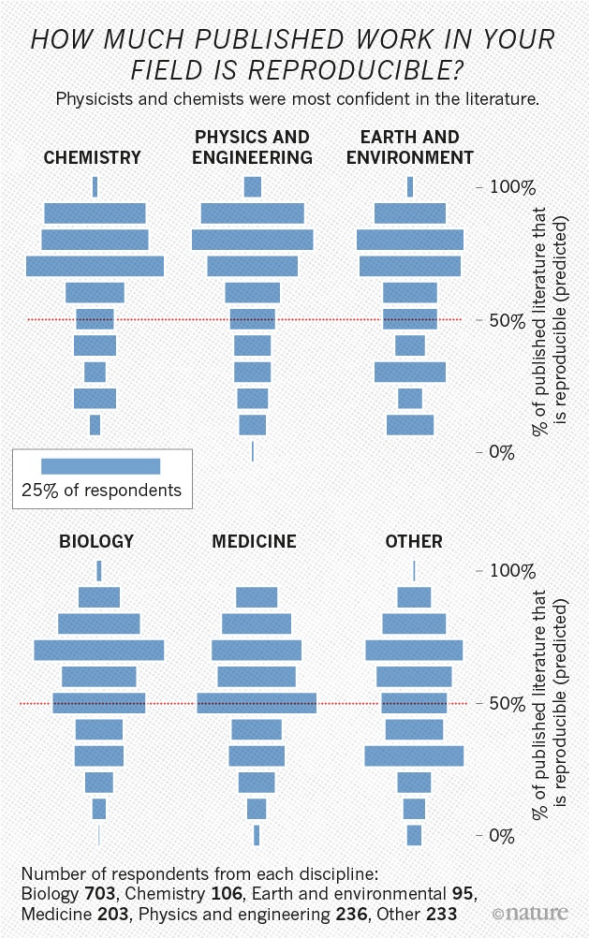
ICT Solutions for Brilliant Minds

The FAIR principles
An Introduction to Research Data Management

Jessica Parland-von Essen

https://orcid.org/0000-0003-4460-3906

9.5.2019



Monya Baker: 1,500 scientists lift the lid on reproducibility. Survey sheds light on the 'crisis' rocking research. Nature 533, 2016. doi:10.1038/533452a

The reproducibility crisis

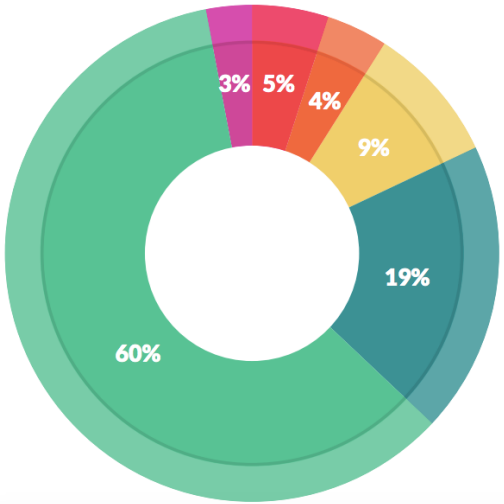
- the virtual absence of replication studies in the published literature in many scientific fields (e.g., Makel, Plucker, & Hegarty 2012),
- widespread failure to reproduce results of published studies in large systematic replication projects (e.g., OSC 2015; Begley & Ellis 2012),
- evidence of publication bias (Fanelli 2010a),
- a high prevalence of "questionable research practices", which inflate the rate of false positives in the literature (Simmons, Nelson, & Simonsohn 2011; John, Loewenstein, & Prelec 2012; Agnoli et al. 2017; Fraser et al. 2018), and
- **the documented lack of transparency and completeness in the reporting of methods, data and analysis in scientific publication (Bakker & Wicherts 2011; Nuijten et al. 2016).**

https://plato.stanford.edu/entries/scientific-reproducibility/

EVIDENCE

Credit: Center for Domestic Preparedness, Federal Emergency Management Agency, U.S. Department of Homeland Security

Working with data: The amount of effort



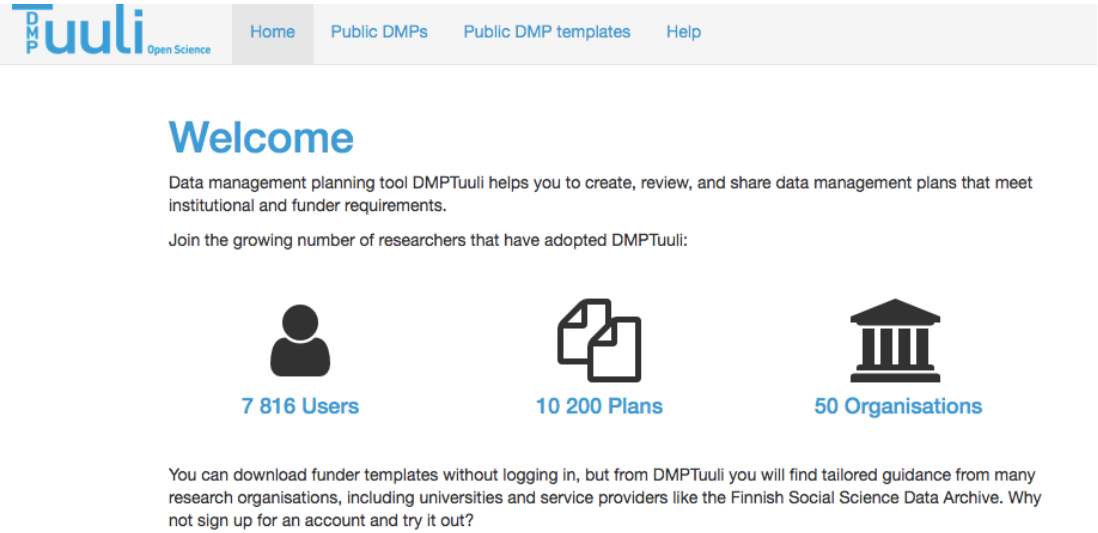
What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

Data Science Report 2016 http://visit.crowdflower.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf

Working with data: Lifecycle management

- Versioning and provenance
- Findability and usability



Data management and data lifecycle

- Think about the data lifecycle from planning to re-use and vice versa
- Ground your data management plan in available skills and good practice in your field
- Make sure it is feasible to implement your plan
- Seek advice and utilize services provided locally, nationally and internationally

Data management: Why do it?

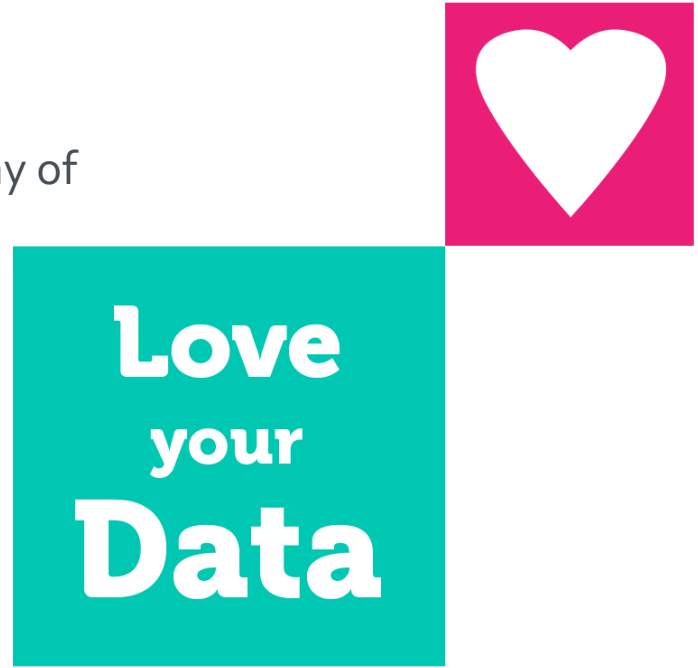
- Make informed decisions and develop procedures
- Ensure data are accurate, complete, reliable and secure
- Avoid duplication, data loss and security breaches
- Stop yourself drowning in irrelevant data
- Share your data for re-use and collaborations
- Write a data paper and publish your data
- Get credit for your data and the work that went into it

Cartoon by Auke Herrema

Make research easier!

Incentives for openness

- Transparency builds trust
- **Research funders requirements** (Horizon, Academy of Finland etc.)
- PSI – public information should be open
- Efficiency and cost effectiveness
- More impact and visibility
- Metadata can be open even if data cannot
- Rights management and documentation improves

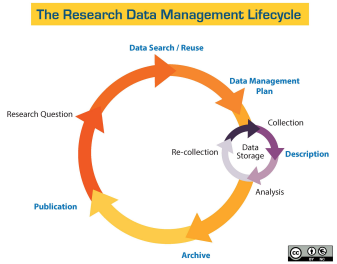


As open as possible, as closed as necessary

- contractual - Saatavuutta rajoitettu sopimuksen perusteella, esim. luottamuksellisen kaupallisen tai teollisen toiminnan perusteella
- personal_data - Saatavuutta rajoitettu luottamuksellisiin henkilötietoihin perustuen
- personal_interest - Saatavuutta rajoitettu tietoja antaneen henkilön etuun tai suojaan perustuen tai esim. luovutussopimuksen perusteella
- national_interest - Saatavuutta rajoitettu kansainvälisiin suhteisiin, yleiseen turvallisuuteen tai kansalliseen puolustukseen perustuen
- judicial - Saatavuutta rajoitettu tuomioistuinkäsittelyn tai oikeudenkäyntiin perustuen
- environmental - Saatavuutta rajoitettu luonnonsuojelun perusteella
- copyright - Saatavuutta rajoitettu tekijäoikeuden perusteella
- cultural - Saatavuutta rajoitettu kulttuuriperinnön tai alkuperäiskansojen suojelun perusteella
- research - Saatavuutta rajoitettu sopimuksen perusteella vain tutkimuskäyttöön
- education - Saatavuutta rajoitettu sopimuksen perusteella opetukseen ja opiskeluun
- other - Saatavuutta rajoitettu muulla perusteella

The FAIR principles

FAIR data is good data



Rutgers lib guide

l.com/ , which provides metadata for institutions identified

ıation, which also comes from Ringgold, but exposes

act matching on tuples of (name,city,region,country). This
) 000 institutions covered: this high matching rate can be
database. Not all Ringgold identifiers are recovered, because
fail for certain names.

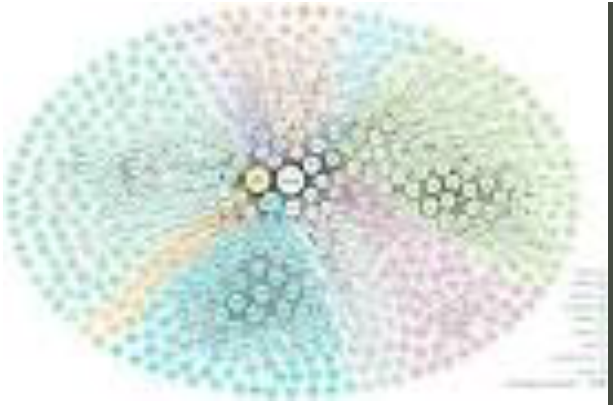
Publication date:
August 24, 2017

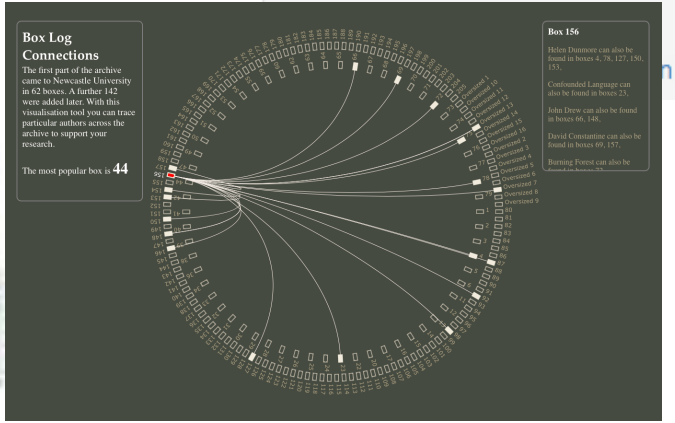
DOI:
DOI 10.5281/zenodo.844869

Keyword(s):
lsnl ringgold

Communities:
Zenodo

n 4.0





Schofield et al (2015) http://digitalhumanities.org:8081/dhq/vol/9/3/000227/000227.html

The FAIR principles for research data

F

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

A

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

I

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

R

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

Wilkinson M et al, "The FAIR Guiding Principles for scientific data management and stewardship". *Scientific Data* (2016/03/15/online). http://dx.doi.org/10.1038/sdata.2016.18

F

1.(Meta)data are assigned a globally unique and persistent identifier

2. Data are described with rich metadata

3. Metadata clearly and explicitly include the identifier of the data it describes

4. (Meta)data are registered or indexed in a searchable resource

A

1. (Meta)data are retrievable by their identifier using a standardized communications protocol

2. Metadata are accessible, even when the data are no longer available

1.(Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

2.(Meta)data use vocabularies that follow FAIR principles

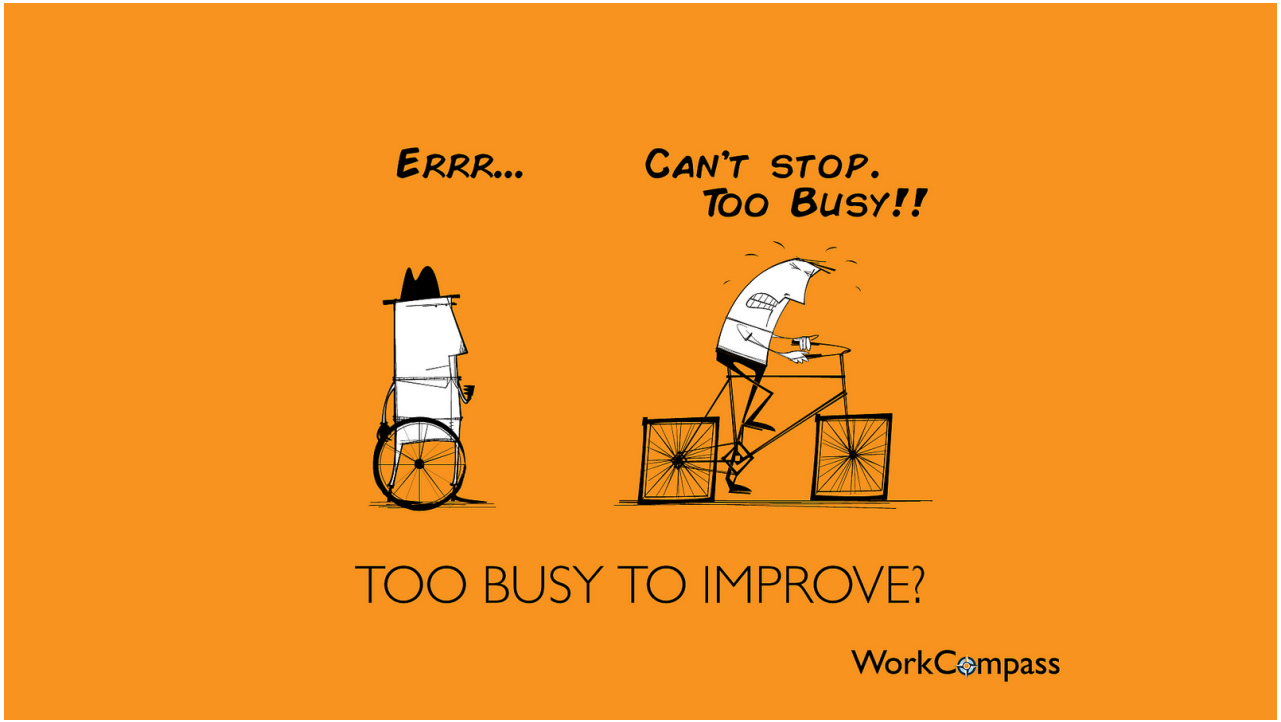
3. (Meta)data include qualified references to other (meta)data

R

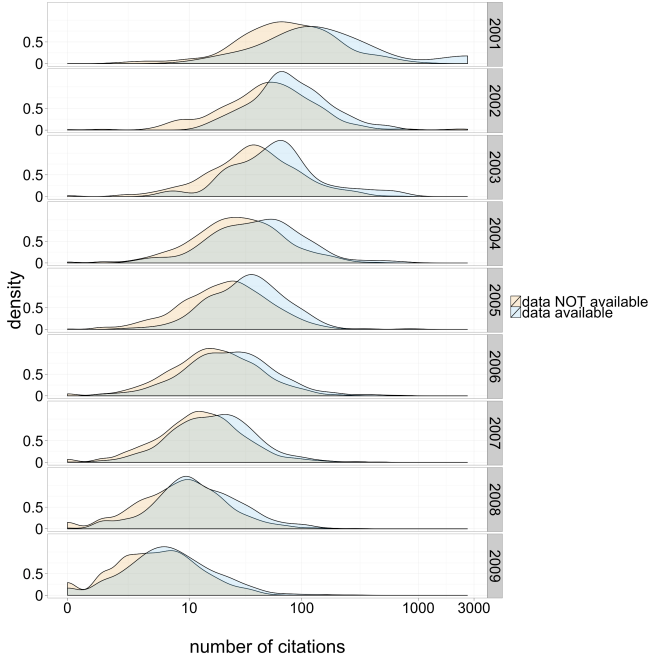
Meta(data) are richly described with a plurality of accurate and relevant attributes

How can a researcher create FAIR data?

1. Write a data management plan and keep it up to date
2. Pay attention to the FAIR principles from the very start
3. Use common data formats
4. Document your data, create rich metadata
5. Use suitable metadata standards
6. Put you data in a data repository
7. Choose a repository that offers persistent identifiers
8. Licence your data
9. Cite data, yours and others!



FAIR data gives impact



Piwowar HA, Vision TJ. (2013) Data reuse and the open data citation advantage. PeerJ 1:e175 https://doi.org/10.7717/peerj.175

F

- Data is described in a relevant catalog with sufficient information
- The dataset has a landing page and a unique and global identifier

A

- The data can be retrieved over the internet
- Versioning and lifecycle are documented
- A tombstone page is available even if the data is deleted

I

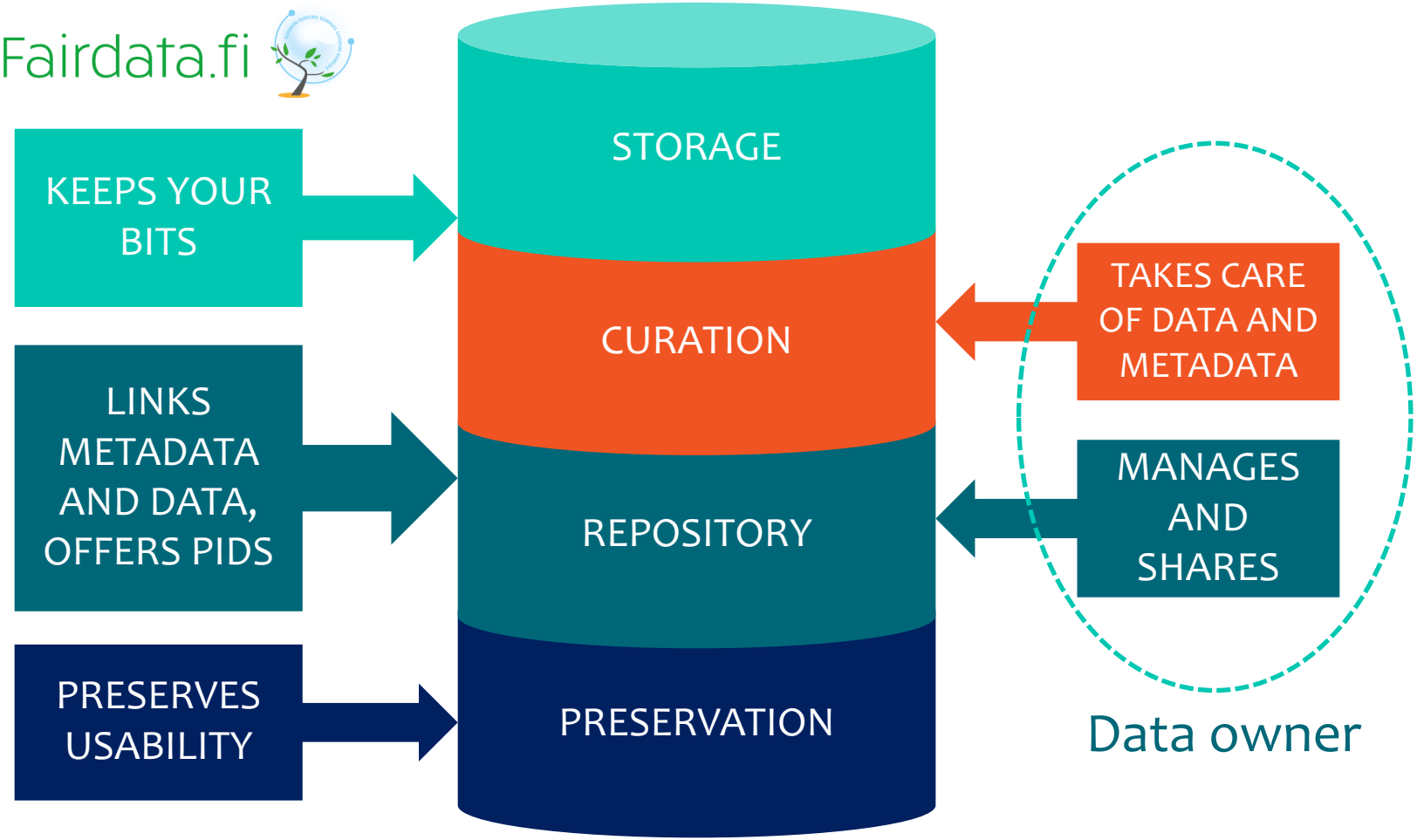
- Use common or at least well documented and preferably open formats

R

- Rights and possible licenses are clearly stated
- Data is well documented and intelligible

Choosing services

Fairdata.fi


KEEPS YOUR BITS
LINKS METADATA AND DATA, OFFERS PIDS
PRESERVES USABILITY
STORAGE
CURATION
REPOSITORY
PRESERVATION
TAKES CARE OF DATA AND METADATA
MANAGES AND SHARES
Data owner

Fixed research data

- For scientific citation
- Unique result of research process
- Versioning
- Important for replication
  and reproducibility
- DOI and ORCID are important
- Models, societal datasets

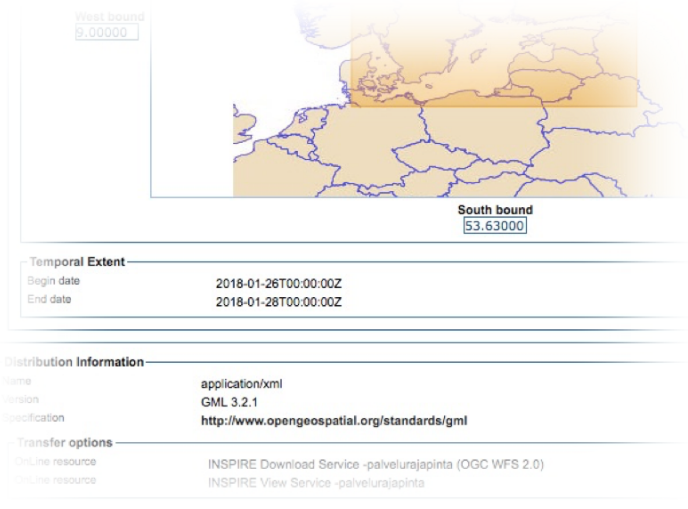
DRIVEN BY doi®

Cumulative Research Data

- On going campaigns
- Consistent format
- Coherent datasets, configurations
- For research
- Data might be added
- If data is retracted or changed a new version with new PID is created
- Best practice http://bit.ly/2Gjrknj

Picture by Tumi-1983 (Own work) [Public domain], via Wikimedia Commons

Dynamic data



- Separate APIs and/or databases
- Might be open data, PSI
- Examples in http://identifiers.org/
- https://www.rd-alliance.org/system/files/documents/RDA-DC-Recommendations_150609.pdf

What defines a good repository?

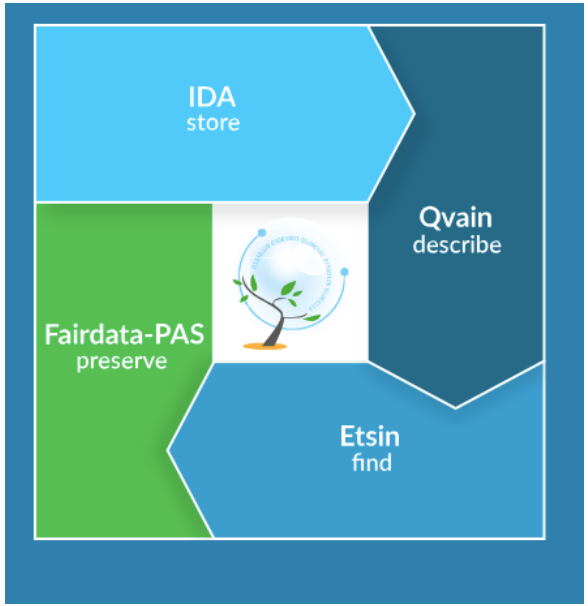
- Established
- Good metadata
- Curation and preservation
- Open and machine readable
- Re3data.org

Type	Recommended	Avoid for data sharing
Tabular data	CSV, TSV, SPSS portable	Excel
Text	Plain text, HTML, RTF PDF/A only if layout matters	Word
Media	Container: MP4, Ogg Codec: Theora, Dirac, FLAC	Quicktime H264
Images	TIFF, JPEG2000, PNG	GIF, JPG
Structured data	XML, RDF	RDBMS

Keeping data safe

- Documentation
- Back up
- Quality control
- Evaluate risks
- Agree on rights

The Fairdata.fi portfolio offers tools for research data management



- Sharing
- Storing
- Publishing
- Creating metadata
- Finding datasets
- Citation

1 What do they do? (1/2)

Produce research data

- Save, rename, organize, share data
- **STORE** the data in immutable state

Fairdata.fi IDA

DESCRIBE the data as a research dataset (add metadata)

Fairdata.fi Qvain

Dataset (or just its metadata) is public, **FINDABLE** and citable (PID, landing page)

Fairdata.fi Etsin

Copy dataset to Fairdata PAS for **DIGITAL PRESERVATION** (requires agreement)

Fairdata.fi PAS

Services, which

- help to make research data and related metadata **Findable, Accessible, Interoperable, Re-usable**
- focus on the later parts of data lifecycle, when data are becoming mature

Funders and research journals recommend or may even require access to research data

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

Digital preservation for research data

- Reliable preservation of digital information for several decades or even centuries
- Hardware, software, and file formats will become outdated, while the information must be preserved
- Active monitoring of information integrity and anticipation of various risks
- Metadata has key role
  - information content
  - provenance information
  - how the content can be used
- Services being built, roles and policies under negotiation between MinEdu and universities
  - https://avointiede.fi/tutkimus-pas

B2 Service Suite

B2DROP
Sync and Exchange Research Data

B2SHARE
Store and Share Research Data

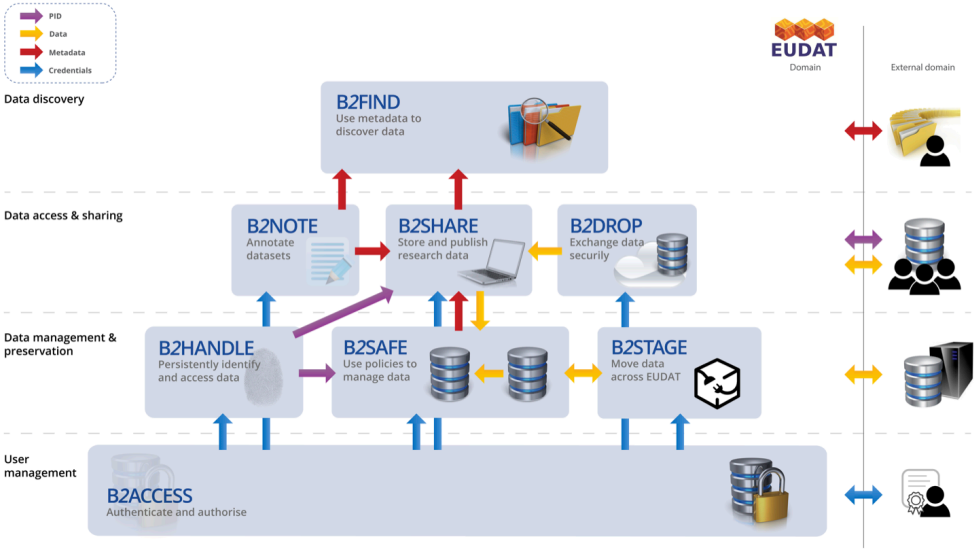
B2SAFE
Replicate Research Data Safely

B2STAGE
Get Data to Computation

B2FIND
Find Research Data

B2HANDLE
Register your Research Data

B2ACCESS
Identity & Authorisation



File management

Organising the files

- Use of Folders – Use folders to group together all the work relevant to your current research/study.
- Use Folder Names that are meaningful –Name the folder(s) so that it relates to your project or area of research.
- Have different folders for any ongoing and completed work.

Naming files

- Use unique file names and the same format throughout your project.
- Keep the file name as short as possible but relevant to research/project. If possible, do not use more than 25 characters.
- Dates in YYYY-MM-DD format allows you to sort your files
- Avoid using special characters such as % & / \ : ; * . ? < > ^! " ()
- Use three digits (or 4 if you have a large number of files) i.e. 001, 002.....201, 202 and not 1, 2, 21, etc.

https://www.rvc.ac.uk/research/about/research-data-management/during-a-project/creating-your-research-data

Naming files

- Use underscores (_) instead of spaces
- If using a personal name in the name of a file give the surname first followed by first name.
- Be careful with personal data when naming files and folders
- Indicate version number by using 'V' followed by the number.

https://www.rvc.ac.uk/research/about/research-data-management/during-a-project/creating-your-research-data

Creating versions

- decide how many versions of a file to keep, which versions to keep, for how long and how to organise versions
- identify milestone versions of files to keep
- uniquely identify files using a systematic naming convention
- record version and status of a file, e.g. draft, interim, final, internal
- record what changes are made to a file when a new version is created

data-archive.ac.uk/create-manage/format/versions

Creating versions

- record relationships between items where needed, e.g. relationship between code and the data file it is run against; between data file and related documentation or metadata; or between multiple files
- track the location of files if they are stored in a variety of locations
- regularly synchronise files in different locations, e.g. using MS SyncToy software
- maintain single master files in a suitable file format to avoid version control problems associated with multiple working versions of files being developed in parallel
- identify a single location for the storage of milestone and master versions of files

data-archive.ac.uk/create-manage/format/versions

Jessica Parland-von Essen

parland@csc.fi
@jpve
@jessicapve

facebook.com/CSCfi

twitter.com/CSCfi

youtube.com/CSCfi

linkedin.com/company/csc---it-center-for-science

github.com/CSCfi