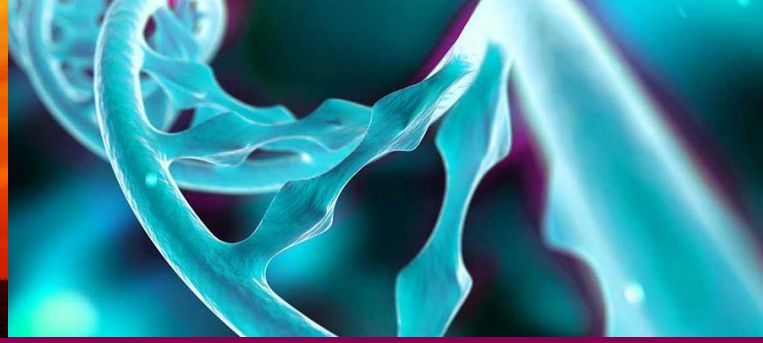


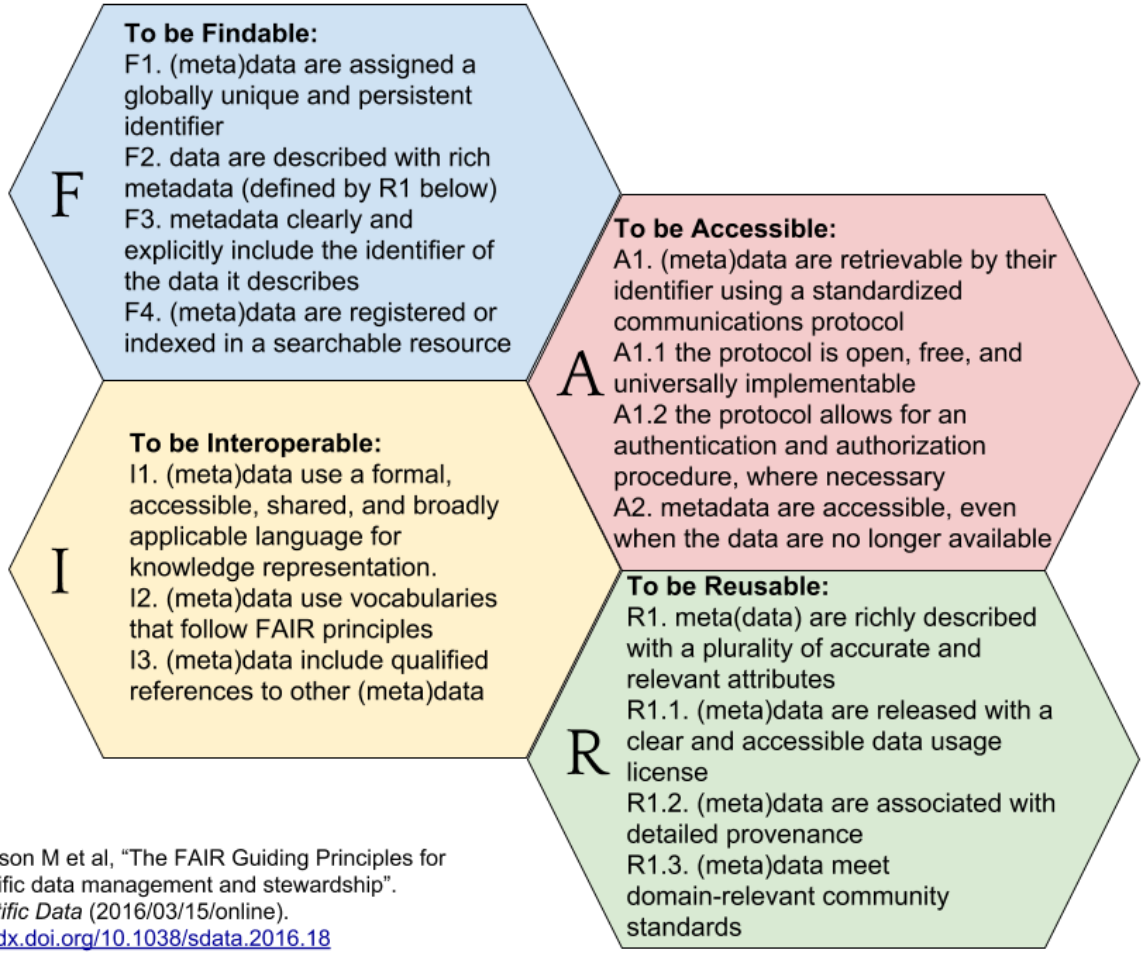


Supporting FAIR data. Categorization of research data as a tool in data management

Jessica Parland-von Essen https://orcid.org/0000-0003-4460-3906, Katja Fält https://orcid.org/0000-0002-6172-5377, Zubair Maalick https://orcid.org/0000-0002-0975-1471, Miika Alonen https://orcid.org/0000-0002-0065-0017, Eduardo Gonzalez https://orcid.org/0000-0003-1400-0995

CSC – Suomalainen tutkimuksen, koulutuksen, kulttuurin ja julkishallinnon ICT-osaamiskeskus

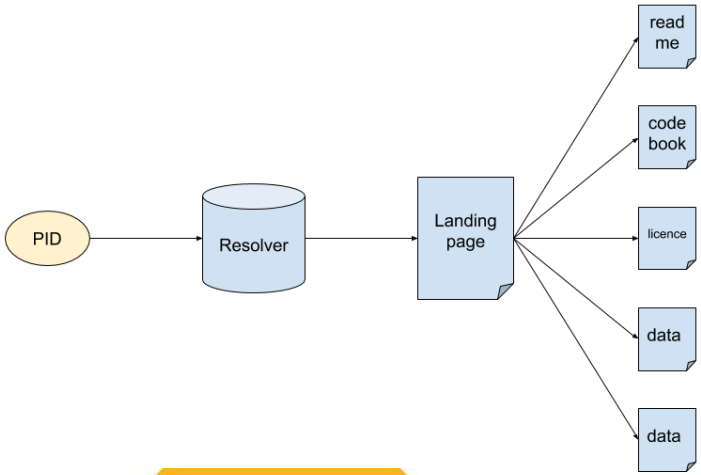
The FAIR principles for research data



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

Persistent identifiers

IMMUTABLE DATASETS



DYNAMIC DATASETS

a) Cite a specific slice or subset (the set of updates to the dataset made during a particular period of time or to a particular area of the dataset).

b) Cite a specific snapshot (a copy of the entire dataset made at a specific time).

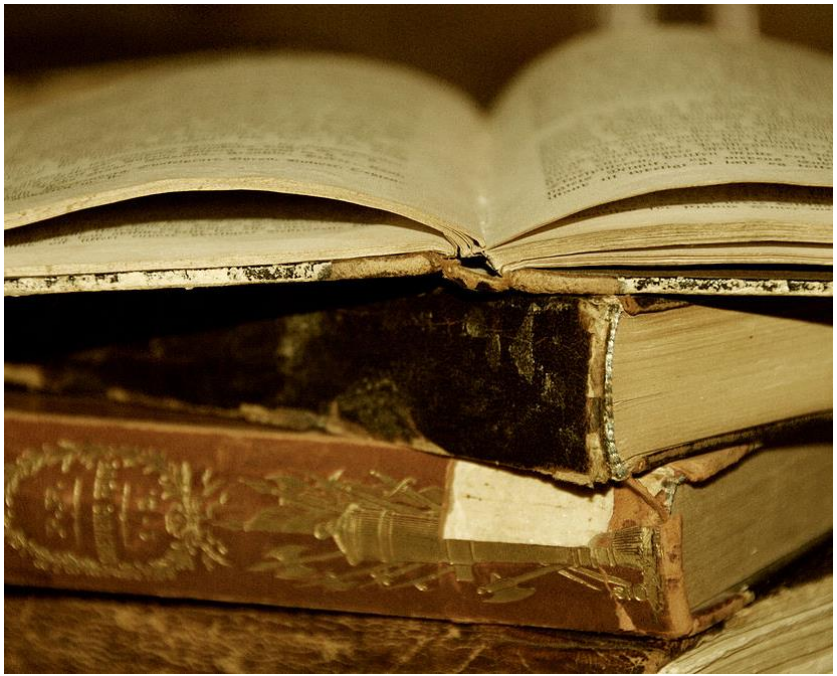
c) Cite the continuously updated dataset, but add Access Date and Time to the citation. (Does not necessarily ensure reproducibility.)

d) Cite a query, time-stamped for re-execution against a versioned database.

Maybe we need to be more specific and find common ground in concepts?

CHUNKING UP RESEARCH DATA

Categorization according to technical properties



- Modality, DCMI types
  - Dublin Core –type of thinking
- Format, DCMI format
  - MIME types
  - Software related
- Language, coding
  - Human interpretation

By Lin Kristensen from New Jersey, USA (Timeless Books) [CC BY 2.0 (https://creativecommons.org/licenses/by/2.0)], via Wikimedia Commons

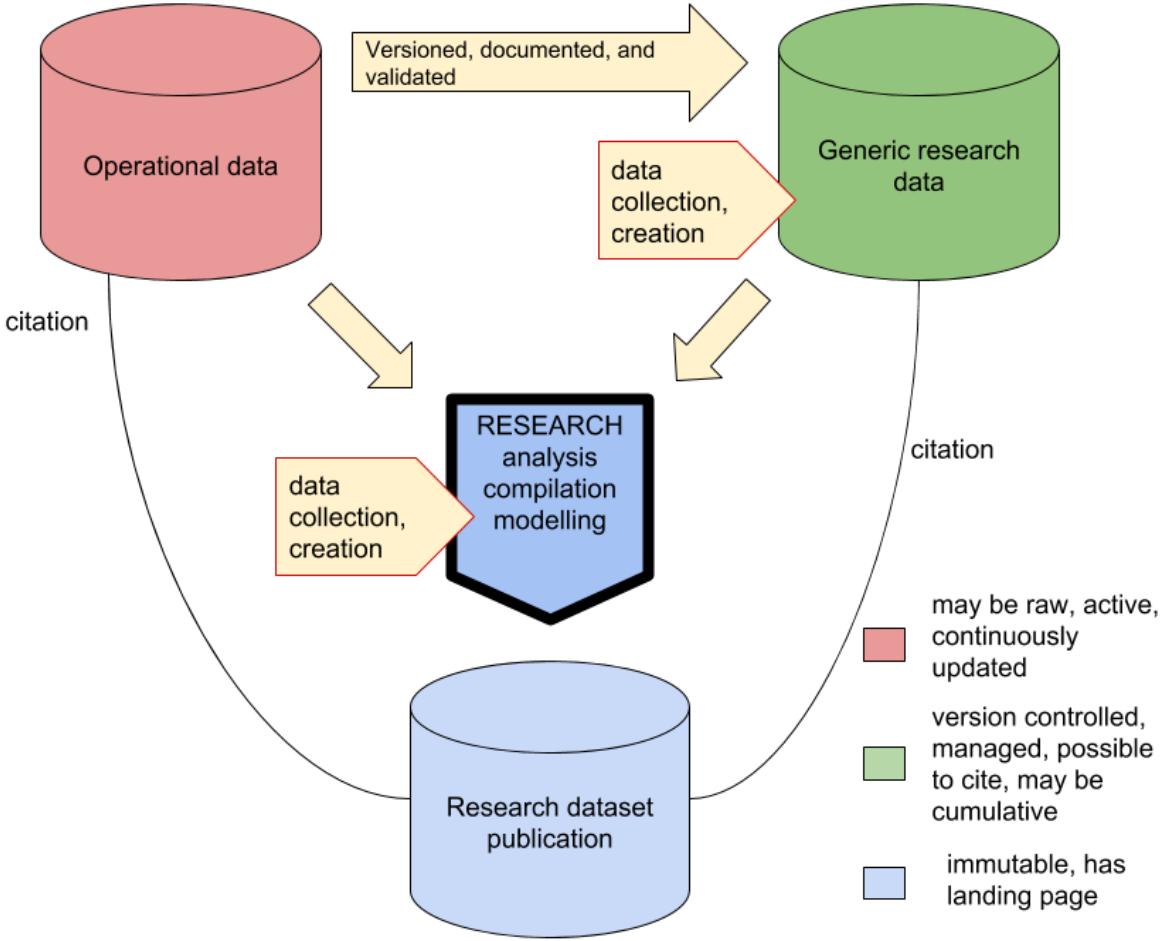
Categorization according to contextual traits

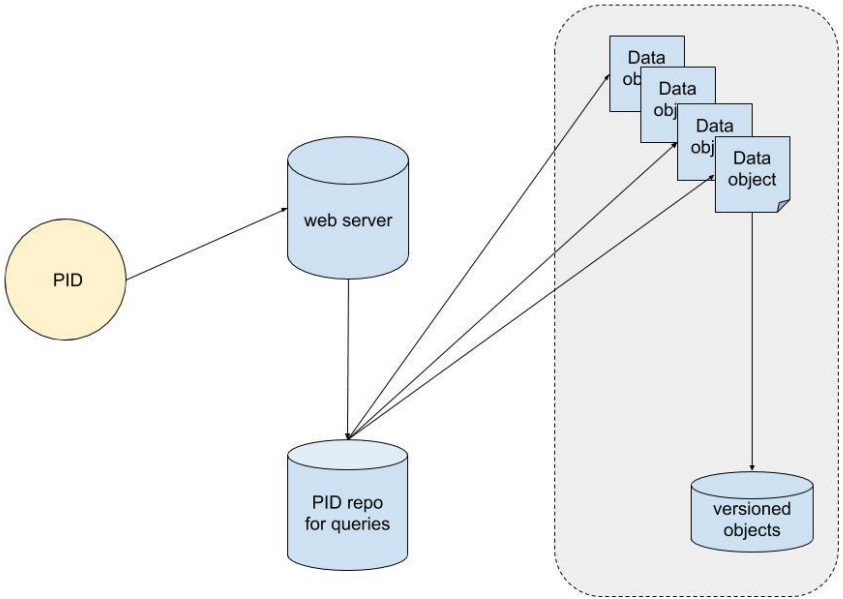
- Origin
  - Observational, experimental, simulation, derived etc
- Use category
  - Source, output, method
- Provenance, lifecycle
  - Primary, secondary, data levels, qualitative, quantitative

By David Monniaux CC-BY-SA-3.0 (http://creativecommons.org/licenses/by-sa/3.0/), from Wikimedia Commons

Categorization according to inherent traits

- Access type (availability)
  - Open data, sensitive data
- Semantic structure
  - Coherence, levels of measurement, groupings, classifications
- Research data type (stability)
  - Generic data, Generic research data, research data publications


Versioned, documented, and validated
Operational data
Generic research data
data collection, creation
citation
RESEARCH analysis compilation modelling
data collection, creation
citation
Research dataset publication
may be raw, active, continuously updated
version controlled, managed, possible to cite, may be cumulative
immutable, has landing page



Dynamic and growing datasets

URN allows use of fragments

Avoid PID inflation

Consider costs and sustainability

Ad hoc creation rather than automatic minting and allocation?

	Operational data	Generic research data	Research dataset
Description	Data for any use, private or government owned, might fall within PSI.	Produced by/with/for researchers, validated, good quality, well documented, might be raw or processed.	Dataset produced for a certain research question Might be highly processed, reuse difficult unless mature field. The main purpose is assessment and reproducibilty.
Format	May be dynamic mature solutions, active or even hot data.	Coherent and well documented formats. Data should be quite stable with versioning. Should be possible to cite and enable reproducible research.	Usually in files, but might also be a database with applications. Citation does not require date. Two-tier resolver for identifier and landing page with metadata available even after data is gone. Might have defined lifespan.
Examples	- weather data - data catalogue - big data from social media	- corpora - time series of experimental or observational data from technical instruments - similar social or clinical surveys	- data paper - data cited in article and published in Zenodo, EUDAT B2Share, other or journal repository

Using research data types …

… makes it easier to describe services

… makes it easier for researchers to plan data life cycle

… makes developing solutions for citation and FAIR data creation and use easier

…makes it easier to describe and manage research data

Jessica PvE parland@csc.fi


facebook.com/CSCfi


twitter.com/CSCfi

youtube.com/CSCfi


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


github.com/CSCfi

Kuvat CSC:n arkisto ja Thinkstock