

DATA PUBLICATIONS ON JÜLICH DATA

... AND DATA PUBLICATIONS IN GENERAL

21 OCTOBER 2026

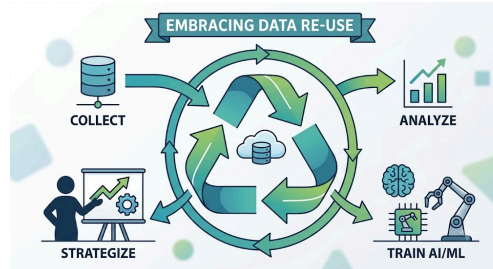
TORSTEN BRONGER

ZENTRALBIBLIOTHEK, FORSCHUNGSZENTRUM JÜLICH

INCENTIVES FOR DATA PUBLICATION



build networks



support re-use culture



advance your career

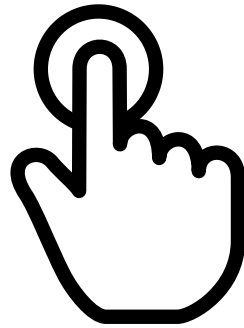


MAKING DATA FAIR

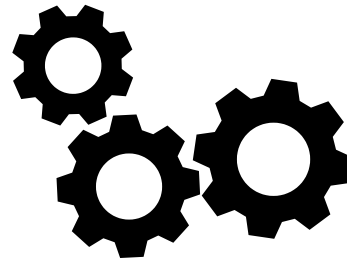
F_{indable}



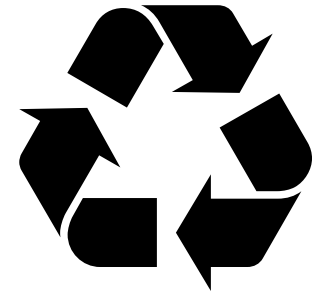
A_{ccessible}



I_{nteroperable}



R_{eusable}



MAKING DATA FAIR

Findable

1. Data is assigned a globally unique and eternally persistent identifier – DOI.
2. Data is described with rich metadata.
3. Data is registered or indexed in a searchable resource.

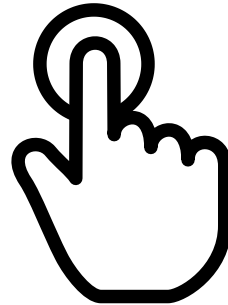


MAKING DATA FAIR

Accessible

1. Data is retrievable by their identifier using a standardised, open communications protocol.
2. Metadata is accessible, even when the data is no longer available.

You may still control who downloads your data!

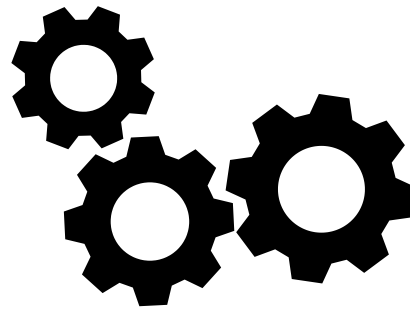


MAKING DATA FAIR

Interoperable

Data uses open data formats.

... or explains its formats in detail.

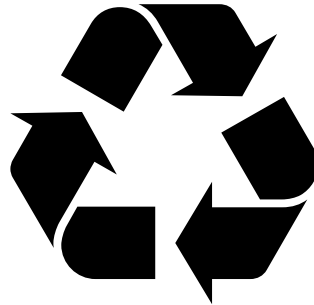


MAKING DATA FAIR

Reusable

Data has a plurality of accurate and relevant attributes:

1. Data is released with a clear and accessible data usage licence.
2. Data is associated with their provenance.



HOW TO MAKE FAIR

Much is already covered if you publish in a decent repository like Jülich DATA.

What you still have to do:

- Document context of the data.
- Describe how to use your data.
- Use file formats that others can ingest easily.

BEYOND FAIR

The FAIR concept is 10 years old. Things that you should do beyond FAIR:

- *Always* allow automatic download of raw data.
- Use controlled vocabulary wherever possible.
- Make your data actionable and interactive with e.g. a Jupyter script.
- Mind granularity.

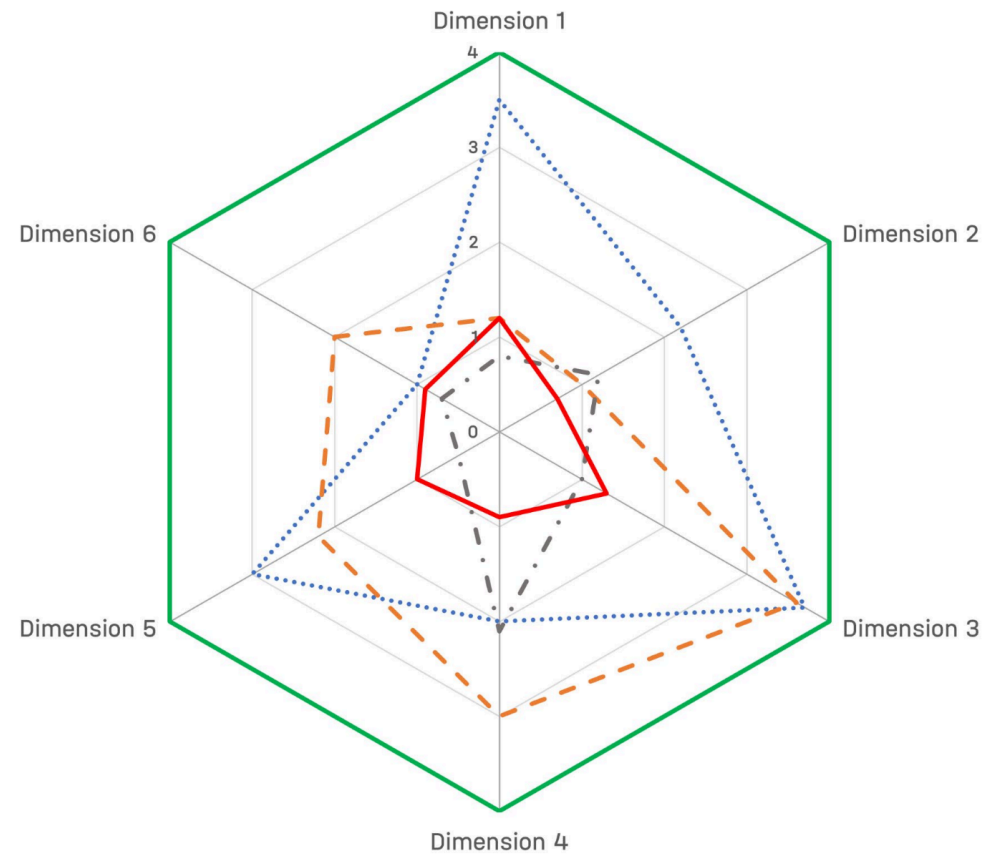
Controlled vocabulary

URLs naming things, e.g. `https://portal.mardi4nfdi.de/entity/Q6673713` means “electrical conductivity”.

The NFDI4ING [Terminology Service](https://terminology.tib.eu/ts/) is a search engine for such URIs.

MAKE DATA COUNT

How the Helmholtz Gemeinschaft will count data and software publications:



A WORD ABOUT GRANULARITY

How big should a data publication be?

- Is every hourly run of your experiment eligible for a data publication?
- Is the complete data of your PhD thesis eligible for one big data publication?

There are no strong rules, however, both is quirky and not recommended. We recommend:

- The size of a data publication should contain as much scientific effort as a typical paper.
- The data contained in a data publication should be self-sufficient, i.e. make sense in its own right.
- A data publication should address one atomic scientific question.

WHAT NEEDS TO BE INCLUDED?

A full data publication is comprised of:

- A **description** that contains all information necessary for an informed expert of the respective field to reproduce and/or reuse your data.
- **Context** and provenance:
 - recipes for preparing the state that was examined
 - properties measured before, or other relevant observations
 - names of things that can be looked-up in other datasets, e.g. sample names or DOIs of papers
- A **licence** for reuse.
- The interoperable **raw data** files, in a sensible directory structure with sensible file names.

GOOD SCIENTIFIC PRACTICE (GSP)

For data publications, the core principles of GSP are:

- Publish all data relevant to a dataset, not only what meets your expectations. Negative results might be valuable, too!
- All the data needs to be archived safely for at least 10 years. This can be delegated to a repository.
- If you are being contacted by someone interested in your raw data, you must have *really* good reason not to provide them with the data. Good reasons might be:
 - You still work on your first publication using the data.
 - Legal issues (e.g. requests from sanctioned countries).
 - Non-disclosure agreements. (Of course, if this means that your results can never be verified, it is not valid scientific output.)
- If you reuse data, cite your sources, no matter which licence they used.

LICENCING

Licencing is important. Besides, it is a low-hanging fruit.

For **data** publications, we recommend two Creative Commons licences:

- CC-0 (totally free reuse), if nobody can hold copyright on the data
- CC-BY (reuse “with attribution”), if there is a copyright holder

Note that in *both* cases, Good Scientific Practice forces a reuser to cite you properly.

Make sure that also your **metadata** has got a licence. Your publication platform may have already taken care of that.

Licencing **software** is a wholly different beast. Basically, you must choose between copyleft (AGPL, GPL, ...) and non-copyleft (MIT, BSD, ...). I recommend AGPL, as it forces reusers to maximal back-contributions.

SOFTWARE PUBLICATIONS

Software publications work very differently, but they are equally important.

- Throwing software on GitHub is not publishing it!
(No persistent identifier; availability is subject to site policies and submitter moods; no enforcement of structured metadata; releases optional, and with opaque URLs.)
- Licence choice is more significant because ownership is clearer than with data publications.
- In the Helmholtz Gemeinschaft, at most one release per year of a certain software is counted as a publication. The exact frequency might yet change, however.

How to do

- Publish every release of your software as a dataset of the special type “software”.
- Use a dedicated software publishing platform like the [Helmholtz Software Directory](https://helmholtz.software/) (<https://helmholtz.software/>).

LOCATIONS FOR DATA PUBLICATION

Where to publish? While your journal article may cite your data, please do not use a commercial Journal's platform for the data publication. Instead:

Jülich DATA



data.fz-juelich.de

Zenodo



zenodo.org

RADAR4ING



radar4ing.scc.kit.edu

NOMAD



nomad-lab.eu

re3data.org



Contact **forschungsdaten@fz-juelich.de**
for assistance
with data and
software
publications on
Jülich DATA!



Mitglied der Helmholtz-Gemeinschaft



JÜLICH DATA

Mitglied der Helmholtz-Gemeinschaft

NEW JÜLICH DATA

A new Jülich DATA is coming in Q4/Q1!

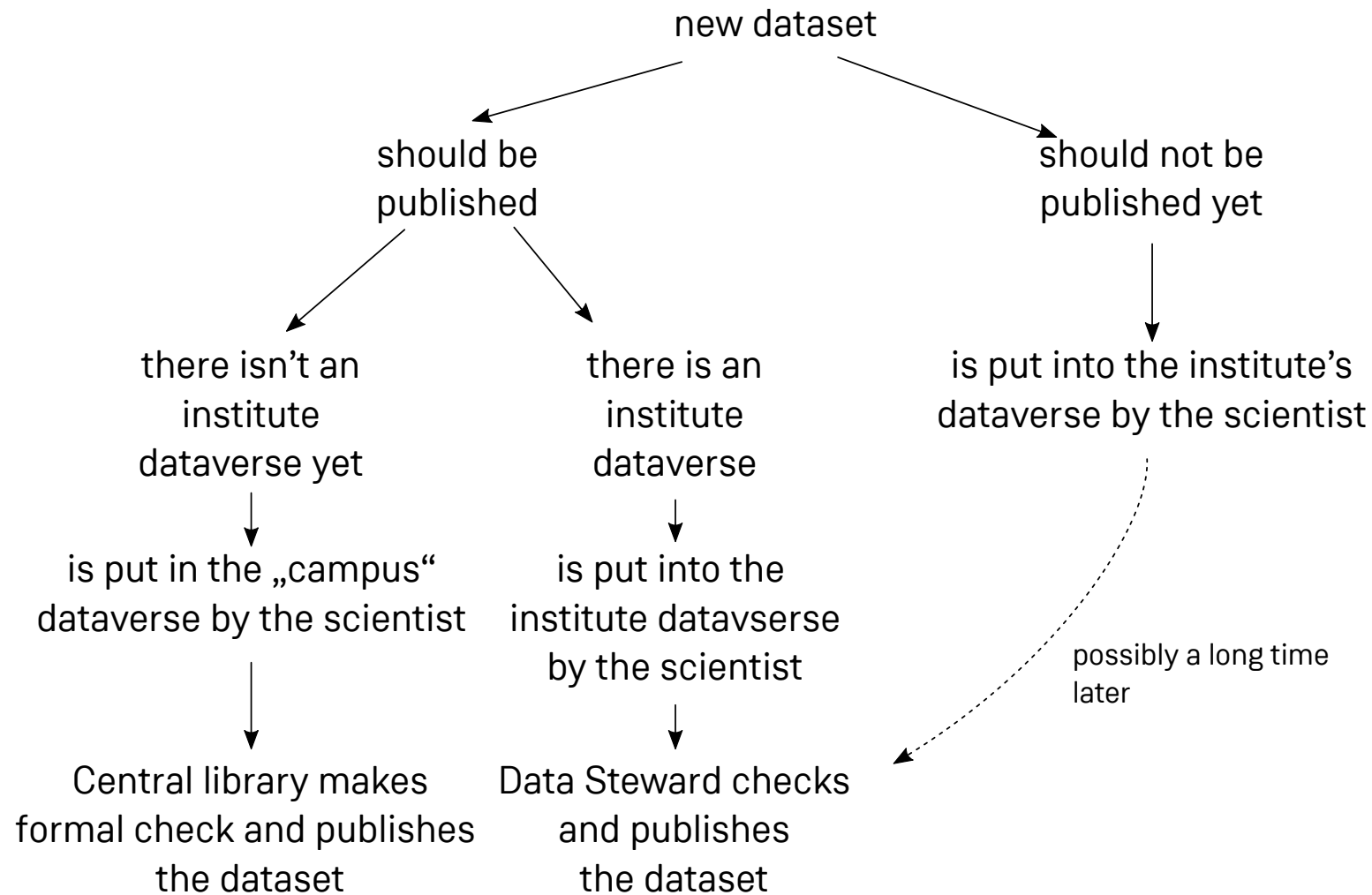
Formerly: Jülich DATA for registering data, i.e. only metadata was stored.

Soon: Jülich DATA also for raw data with 1 PB storage.

- Upgrade of underlying software “Dataverse”:
 - Software publications can be flagged explicitly.
 - API access much better for your scripting.
- Monitoring for higher availability.



WORKFLOW

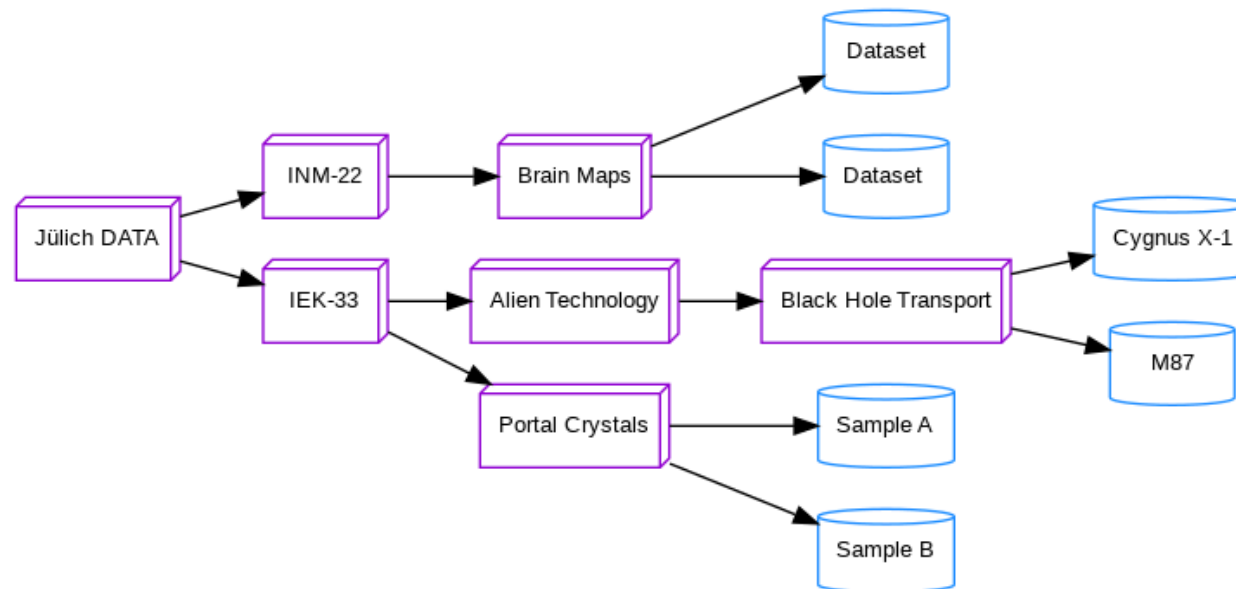


DATAVERSES

The structure of Jülich DATA is similar to a filesystem:

- directories $\hat{=}$ dataverses
- files $\hat{=}$ datasets

Top-level dataverses correspond to institutes.



STORAGE REQUESTS OF INSTITUTES

„Pen's parade“ of storage

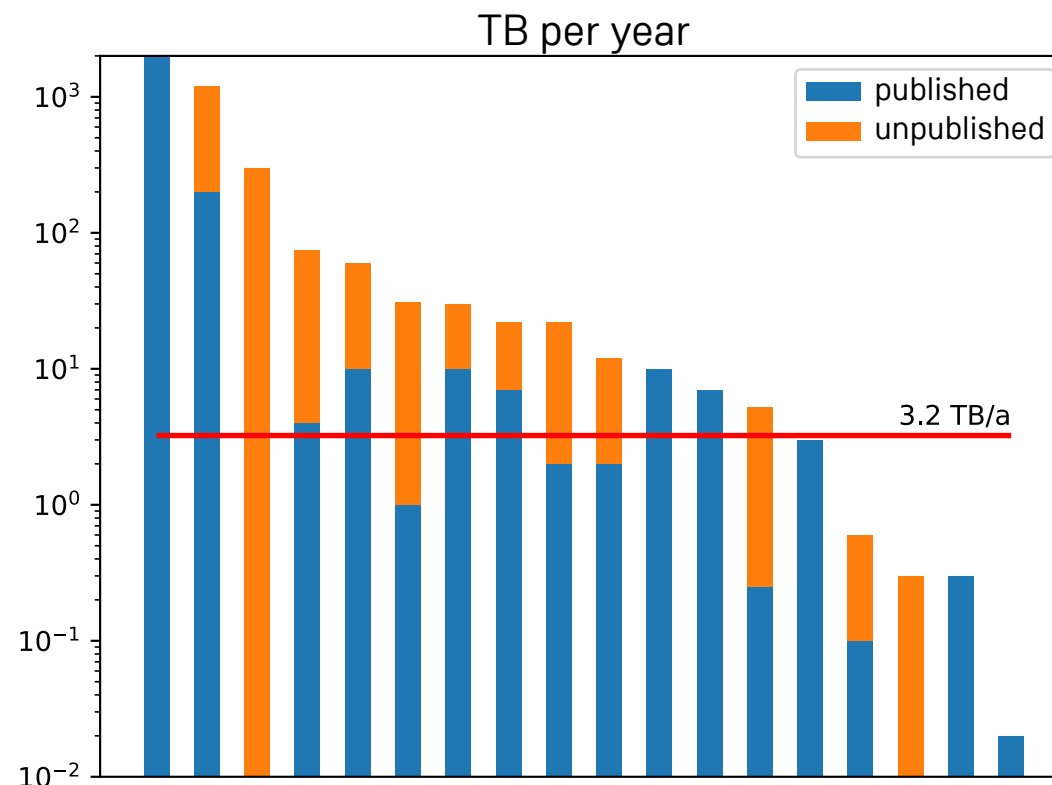
Every column is an institute sorted by requested storage amount.

“unpublished” refers to data that is supposed to be deposited *centrally* in spite of being not published.

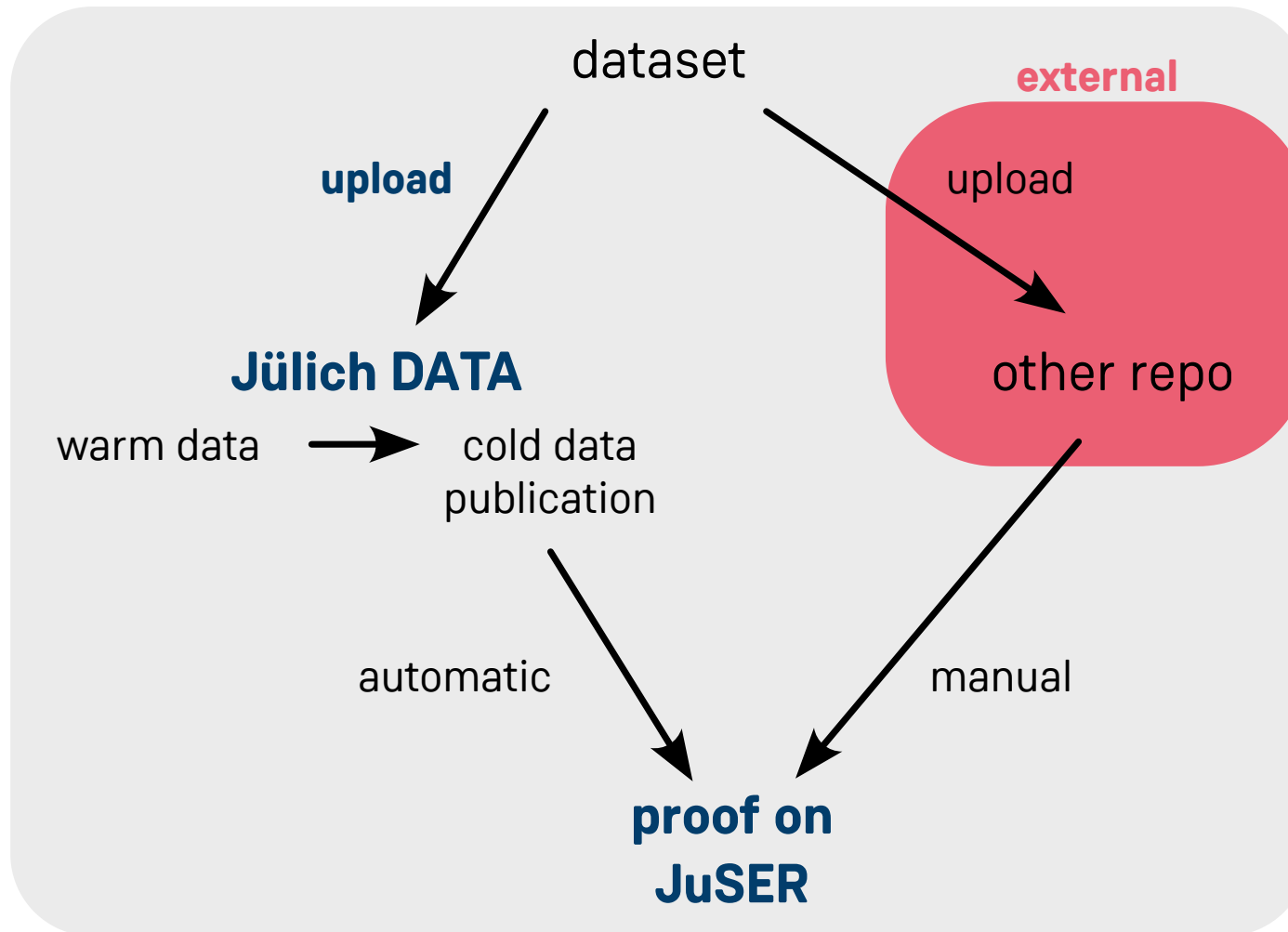
Red line: geometric mean of requests only for *published* data. Extrapolated for all scientific institutes: **220 TB/a**

If Jülich DATA storage is supposed to be sufficient for 3 years: **0,8 PB**

For 5 years: **2 PB**



AUTOMATIC REGISTRATION ON JUSER



TO GET INVOLVED

- If your institute does not have a Data Steward yet – please get in touch with us!
- If you are interested in contributing to the curation rules, size limits etc. – please get in touch with us!
- If your institute or group wants to attach data storage to Jülich DATA for easier data sharing with others – please get in touch with us!

forschungsdaten@fz-juelich.de



THANK YOU

Torsten Bronger

Zentralbibliothek, Forschungszentrum Jülich

t.bronger@fz-juelich.de (<mailto:t.bronger@fz-juelich.de>)

