Friedrich-Alexander-Universität
Erlangen-Nürnberg

FAU

Using the Zenodo document and data repository

Zenodo*

History:

- developed under **OpenAIRE** programme (EU initiative (FP8) to support open access)
- operated by CERN
- launched in 2013 as a document repository; since 2015 also for datasets

* Zenodotus (gr.: Ζηνόδοτος): first head librarian of the Library of Alexandria

Zenodo

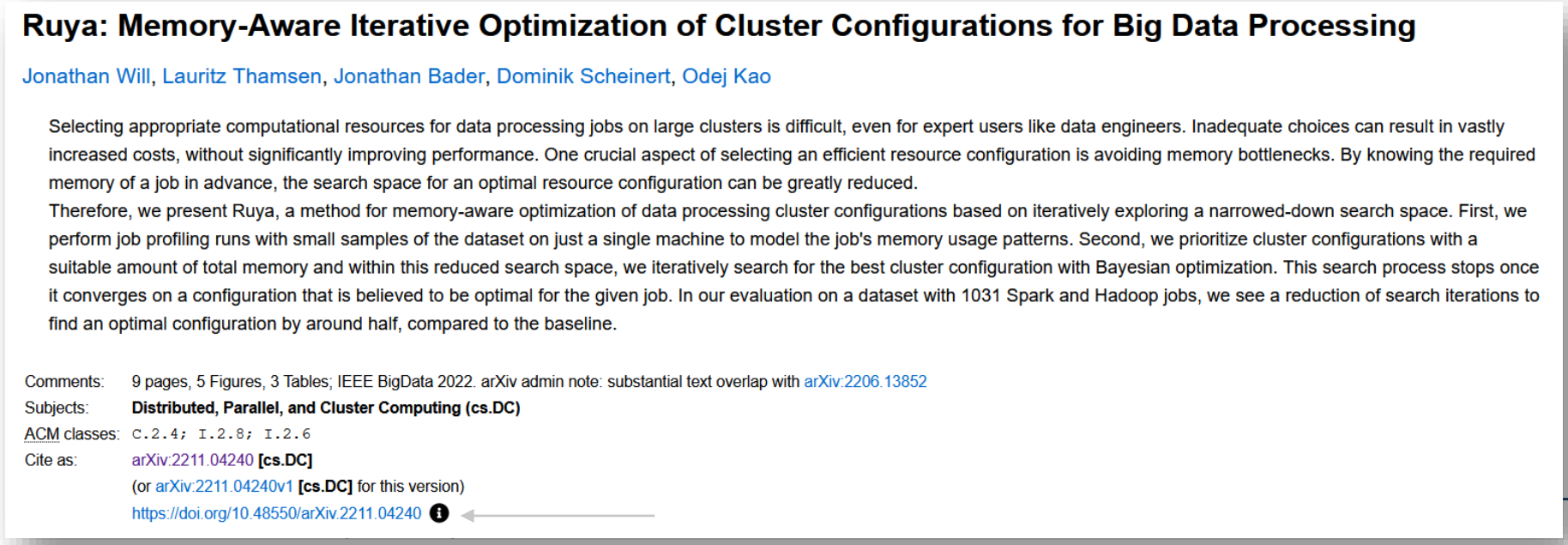
Technical Information:

- primarily hosted in the CERN data centre infrastructure (OpenStack based)
- files stored on CERN EOS system
- metadata in PostgreSQL database
- uses a modified version of the **Invenio software**
  - originally developed by CERN; now an open source (MIT licence) project with multiple contributing institutions / persons
  - „modules“ for generic repositories, integrated library systems & RDM repositories (InvenioRDM = basically Zenodo)
  - also invenio-based: Inspire-hep, B2share, Caltech Library Catalog, …

Zenodo

Features:

- free-of-charge
- size limit on single deposit: 50 GB
- deposits can by assigned to „communities“
- DOI service (including concept DOI and version DOI) + versioning of files



Ruya: Memory-Aware Iterative Optimization of Cluster Configurations for Big Data Processing

Jonathan Will, Lauritz Thamsen, Jonathan Bader, Dominik Scheinert, Odej Kao

Selecting appropriate computational resources for data processing jobs on large clusters is difficult, even for expert users like data engineers. Inadequate choices can result in vastly increased costs, without significantly improving performance. One crucial aspect of selecting an efficient resource configuration is avoiding memory bottlenecks. By knowing the required memory of a job in advance, the search space for an optimal resource configuration can be greatly reduced.
Therefore, we present Ruya, a method for memory-aware optimization of data processing cluster configurations based on iteratively exploring a narrowed-down search space. First, we perform job profiling runs with small samples of the dataset on just a single machine to model the job's memory usage patterns. Second, we prioritize cluster configurations with a suitable amount of total memory and within this reduced search space, we iteratively search for the best cluster configuration with Bayesian optimization. This search process stops once it converges on a configuration that is believed to be optimal for the given job. In our evaluation on a dataset with 1031 Spark and Hadoop jobs, we see a reduction of search iterations to find an optimal configuration by around half, compared to the baseline.

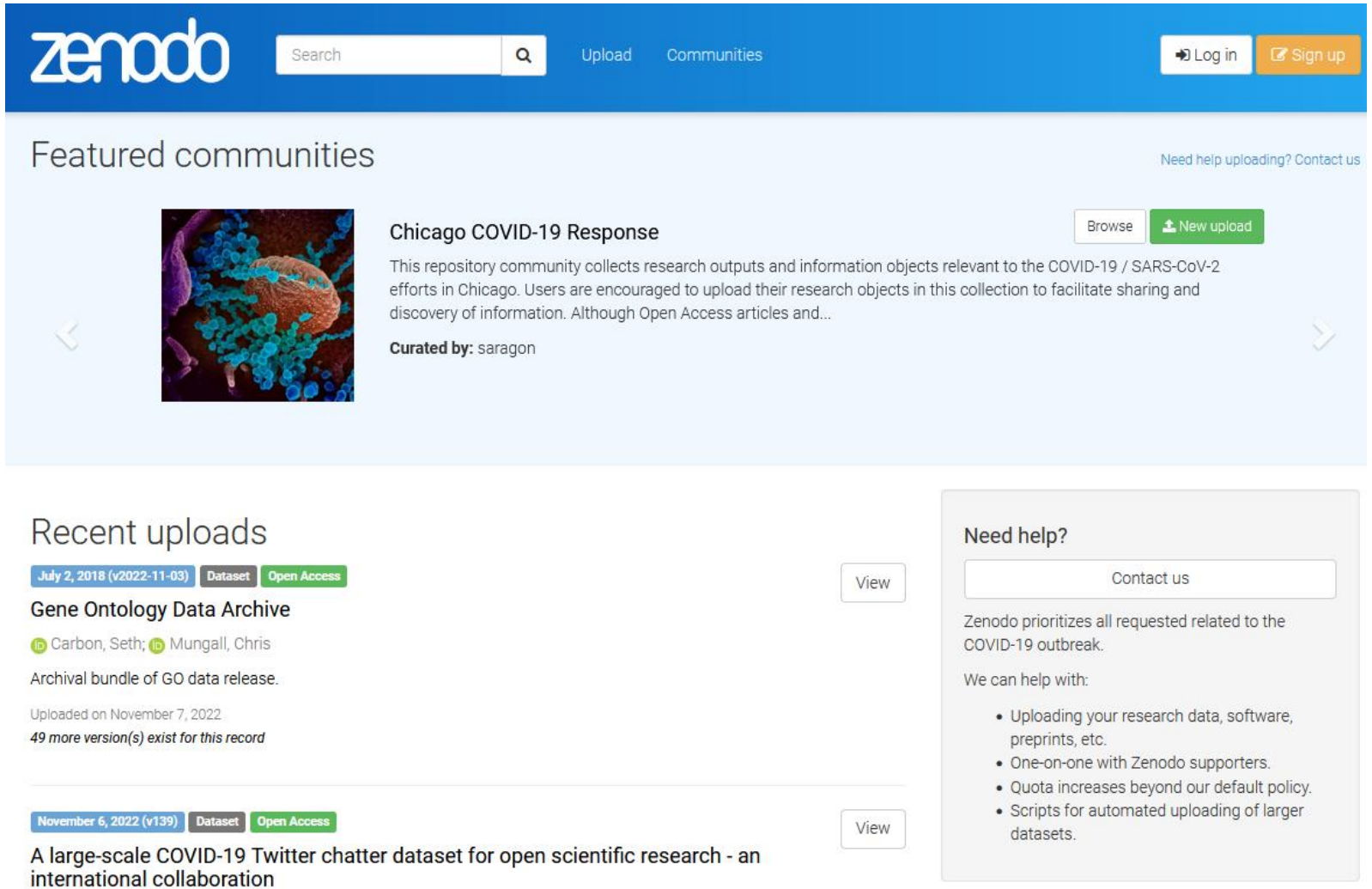
Comments: 9 pages, 5 Figures, 3 Tables; IEEE BigData 2022. arXiv admin note: substantial text overlap with arXiv:2206.13852
Subjects: **Distributed, Parallel, and Cluster Computing (cs.DC)**
ACM classes: C.2.4; I.2.8; I.2.6
Cite as: arXiv:2211.04240 **[cs.DC]**
(or arXiv:2211.04240v1 **[cs.DC]** for this version)
https://doi.org/10.48550/arXiv.2211.04240

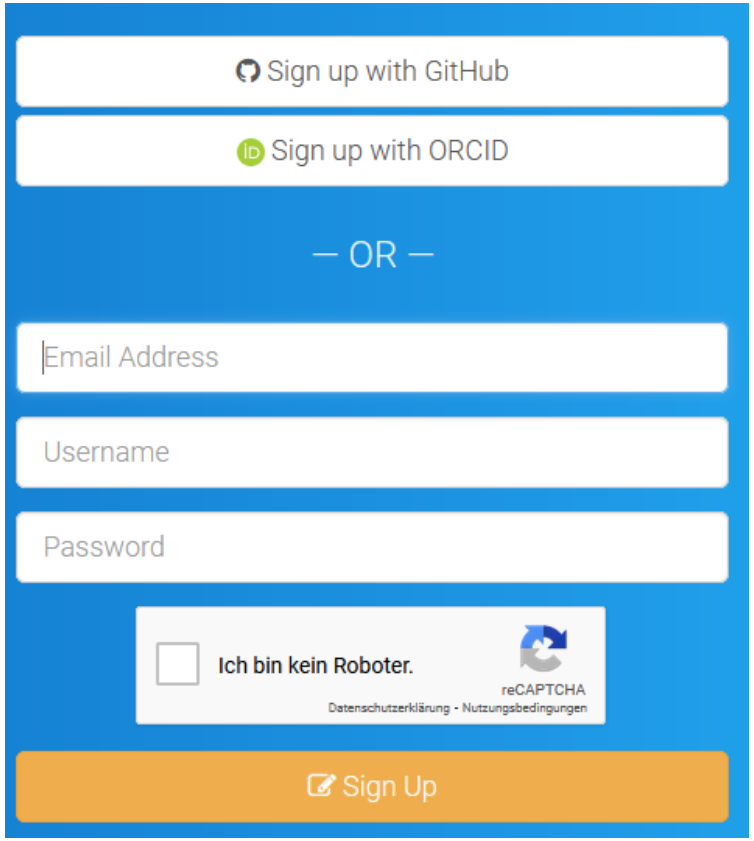
Zenodo

Features:

- free-of-charge
- size limit on single deposit: 50 GB
- deposits can by assigned to „communities“
- DOI service (including concept DOI and version DOI) + versioning of files
- non-static metadata
- Different access levels: open access, restricted, closed (metadata is always CC0→ FAIR complient)
- GitHub integration
- metadata deposited in OpenAIRE (mandatory for EU-funded projects)
- no mandatory file formats / types
- user is responsible for upload (copyright, data protection …)

Zenodo – in practice



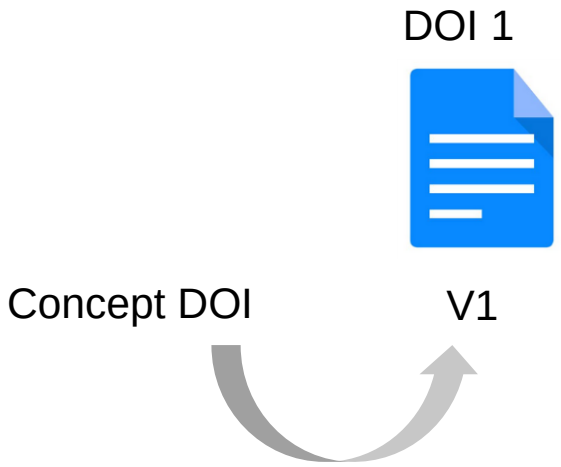


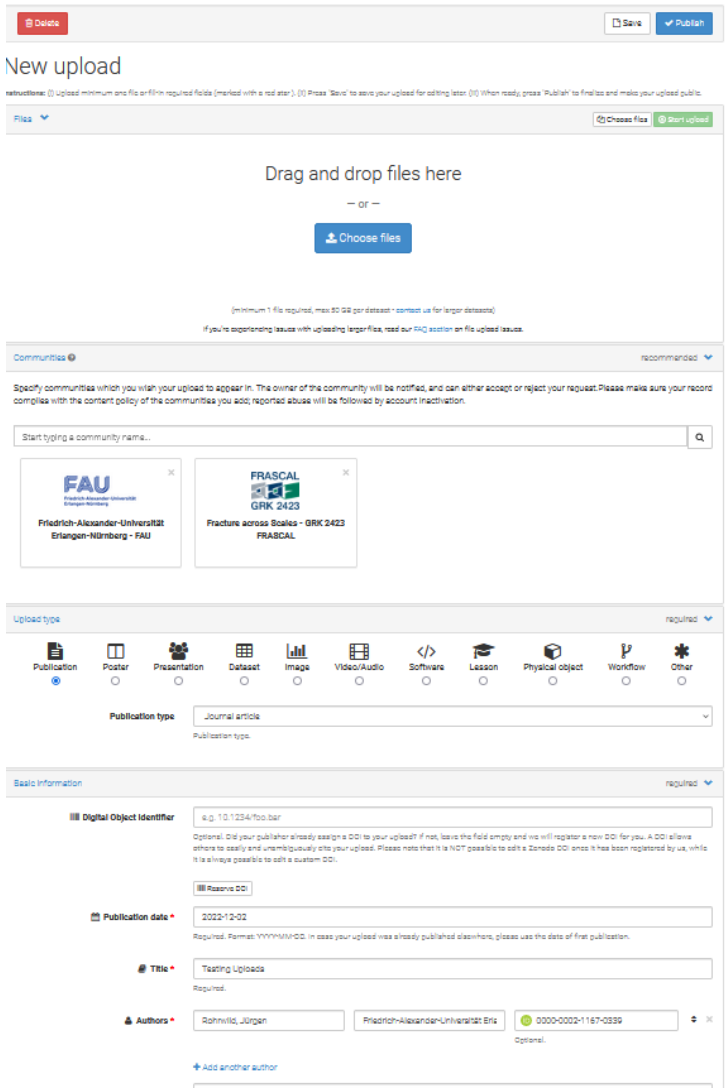
Zenodo – in practice

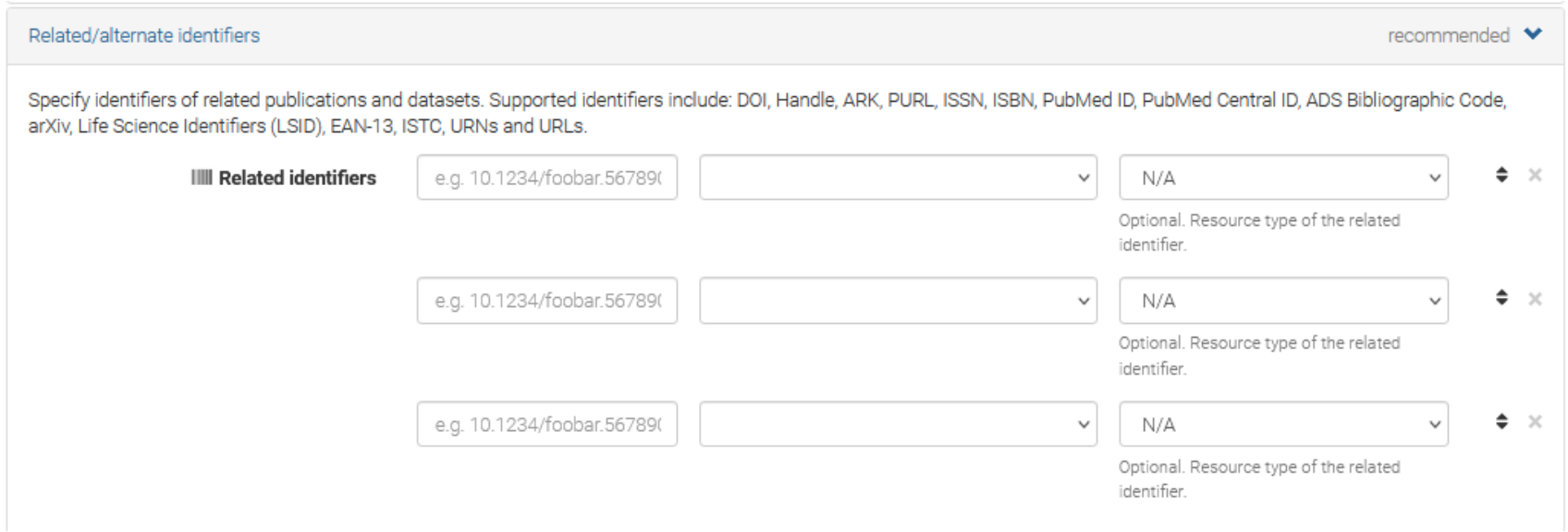
- To sign up:
  - basically any email address → works with a shared mail box etc (useful for managing a community)
  - „full“ functionality after 14 days

Zenodo - Upload

- Metadata and data can be uploaded / entered in „staging mode“ → can be freely changed and saved
- Uploads are tied to the account of the uploader (= only account that can edit / correct metadata)
- After publishing:
  - metadata can still be changed and corrected (e.g. add ORCID, fix a typo in the desription, add a new publication where the dataset / software/ … was used, etc)
  - data is permatently fixed → changing data creates a **new version** (changes are saved incrementally) with a new DOI
  - Zeondo provides a so-called concept DOI



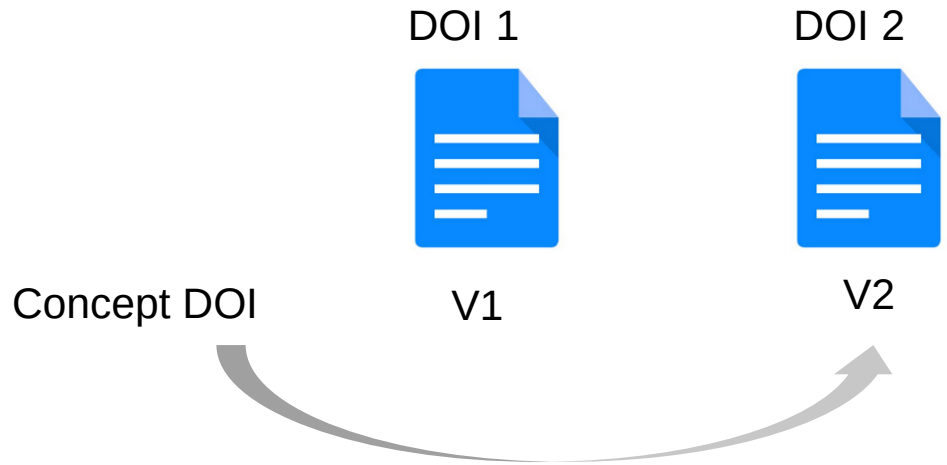


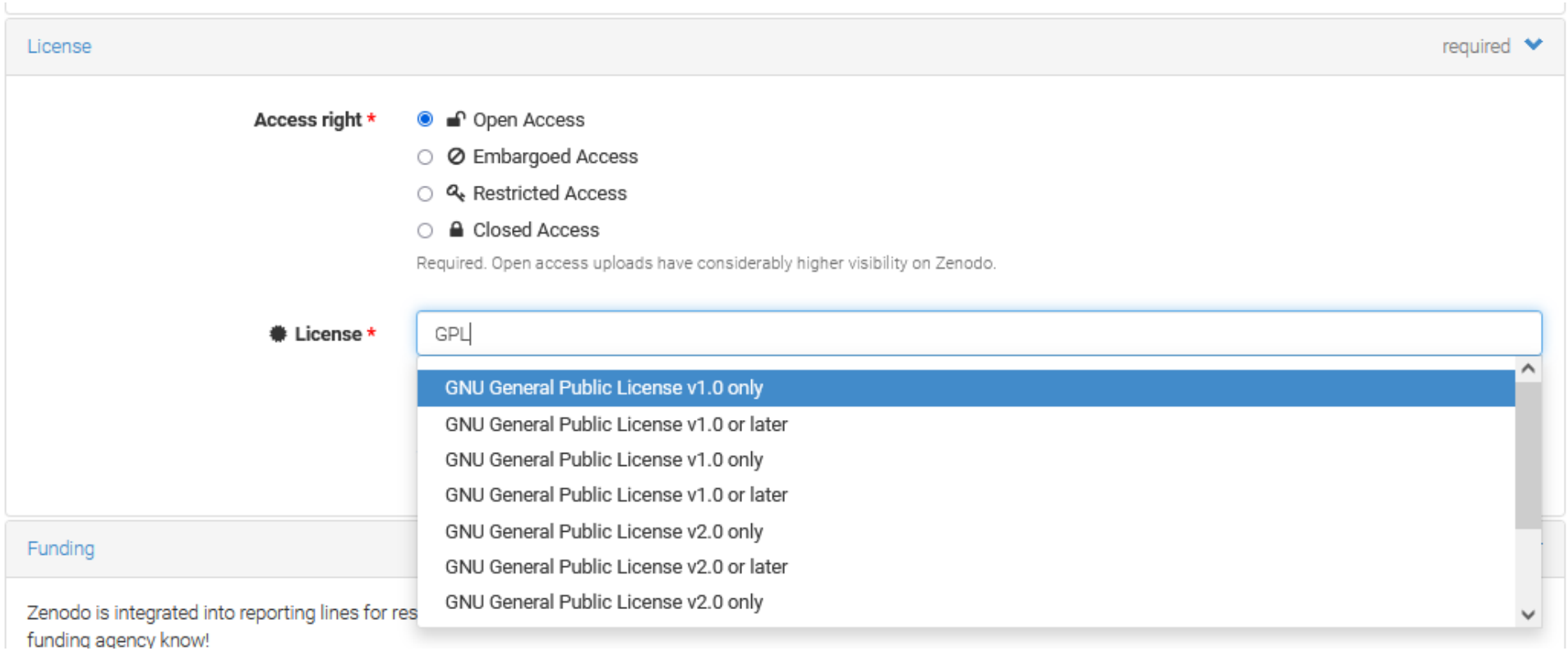


Metadata close to DataCite standard

Zenodo - Upload

- Metadata and data can be uploaded / entered in „staging mode“ → can be freely changed and saved
- Uploads are tied to the account of the uploader (= only account that can edit / correct metadata)
- After publishing:
  - metadata can still be changed and corrected (e.g. add ORCID, fix a typo in the desription, add a new publication where the dataset / software/ … was used, etc)
  - data is permatently fixed → changing data creates a **new version** (changes are saved incrementally) with a new DOI
  - Zeondo provides a so-called concept DOI





- broad set of licence options
- four access options

Communities

Friedrich-Alexander-Universität Erlangen-Nürnberg - FAU

The Friedrich-Alexander University Erlangen-Nürnberg (FAU) is a strong research university with an international perspective and one of the largest universities in Germany, with around 40,000 students, 260 degree programmes, 4,000 academic staff (including over 580 professors), 200 million euros third-party funding, and 500 partnerships with universities all over the world. Teaching at the University is closely linked to research and focuses on training students in both theory and practice to enable them to think critically and work independently. The research itself also strikes the perfect balance between a theoretical approach and practical application. Only fast, direct and ideally free access to academic publications and primary research data unlocks the full potential of this research environment. Visit https://www.fau.de/ to learn more about the FAU.

Zenodo FAU community collection

Feel free to use this Zenodo collection FAU as a central electronic archiving and publication platform for research data which can´t be archived or published on OPUS FAU (http://opus4.kobv.de/opus4-fau/home), RRZE basis storage (https://www.rrze.fau.de/files/2017/06/Betreuungsvereinbarung_Basis-Storage.pdf) or institutional websites. Qualified scientific research data of all members of FAU may be published here free of charge. Corresponding full texts to the research data shall be published on OPUS FAU or disciplinary repositories. Thus all publications and research data sets are permanently available to the global public and are searchable and citable via catalogues and search engines.

For the creation of data management plans the tools RDMO (https://rdmorganiser.github.io/) or DMPonline (https://dmponline.dcc.ac.uk/) can be useful. For further information visit e.g. the Digital Curation Centre (DDC) (http://www.dcc.ac.uk/resources/data-management-plans). FAU also offers advice on legal and organizational matters to all researchers who would like to publish their research in open access media, via its Open Access Policy (urn:nbn:de:bvb:29-opus4-68651).

Only free and direct access to academic publications and primary research data unlocks the full potential of a digital research environment, academic networks and integrated research databases. The University thus wants to promote open access in the long term and has established an open access publishing fund to cover these costs (https://ub.fau.de/en/writing-publishing/open-access/).

Since this community is operated by the 'Referat Open Access', opening and making the data available (i.e., maximum restricted access, please no closed access) is desired.

Subject-based research data repositories

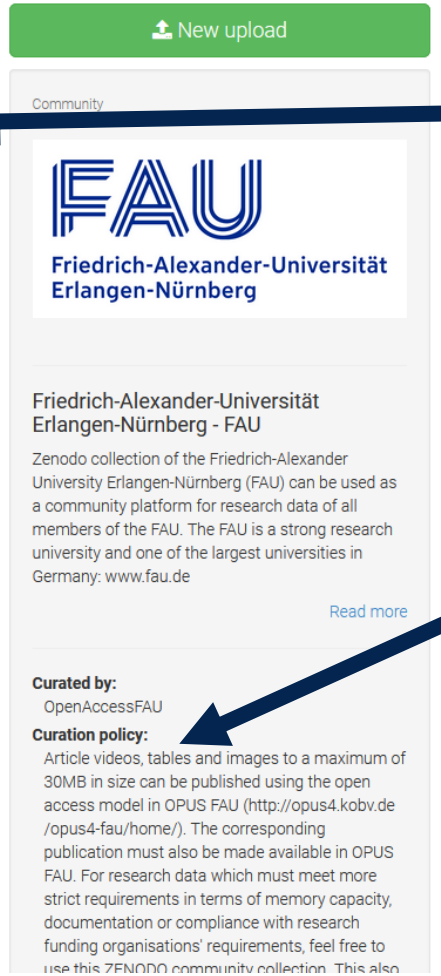
Other research data repositories can be found sorted by subjects on the web-based directory of the project re3data.org (Registry of Research Data Repositories, http://www.re3data.org/browse/by-subject/).

Data citation standards

Using the citation standard offers proper recognition to authors as well as permanent identification through the use of global, persistent identifiers in place of URLs. Use of universal numerical fingerprints (UNFs) guarantees to the scholarly community that future researchers will be able to verify that data retrieved is identical to that used in a publication decades earlier, even if it has changed storage media, operating systems, hardware, and statistical program format.

Following are two authentic examples of replication data citations:

From International Studies Quarterly, King and Zeng, 2006, p. 209:

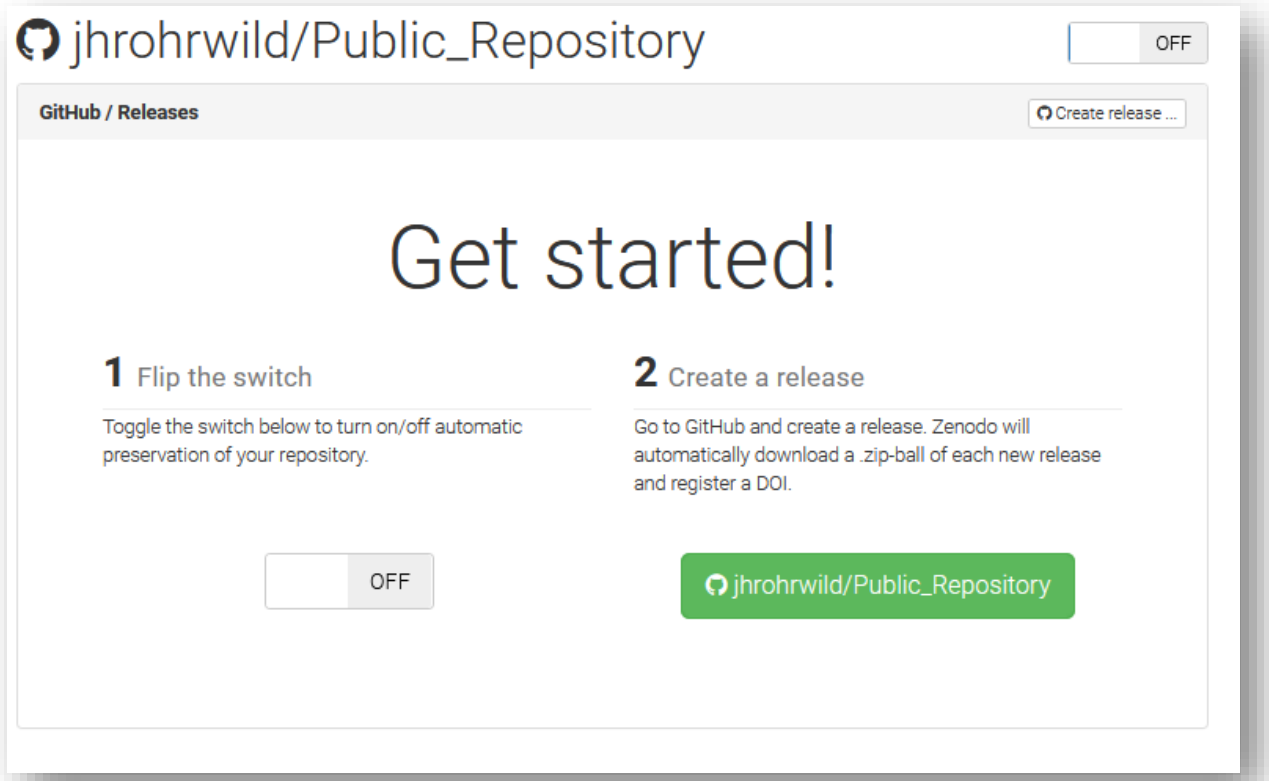


Long description what the community is about.

Curation policy:

What documents are accepted, who decides etc.

Connecting Zenodo with GitHib



Some advice

- publications are tied to the account that uploaded the file
  - get messy if someone changes affiliations / jobs and the metadata has to be fixed

    → consider a team account with a shared mailbox (ub-zenodo@fau.de)
- Communities are nice to present the collected publications (but accounts need to be 14 days old to be able to found a community)
- Zenodo is very user-friendly when it comes to metadata, but you cannot change the data files (or delete them) once published
- Linking Zenodo and GitHub is useful:
  - profit from minted DOI and persistent citability
  - all advantages of GitHub
  - recommended by DFG for published reseach software in mathematics