

FAIR research data integration in CRIS at FAU Erlangen-Nürnberg

Marcus Walther^{1*} and Michael Wagner¹

¹ FAU Competence Unit for Research Data and Information, Friedrich Alexander Universität Erlangen Nürnberg

* Corresponding author: marcus.walther@fau.de; Martensstr. 3, 91058 Erlangen

1 Introduction

Although research data is an important research outcome and there are broadly used meta data standards like the DataCite schema, handling such information in research information systems is still a challenge. We present an implementation for an integrated workflow towards a maximal reusability of data-associated meta information.

1.1 FAU CRIS

Friedrich-Alexander University Erlangen-Nürnberg (FAU) uses Converis [1] as research information system (FAU CRIS, <https://cris.fau.de/>). FAU CRIS covers standard meta data models for research information, like *CERIF* [2] and *Kerndatensatz Forschung* [3]. It has been constantly extended in the last years, for example, with automated affiliation processing [4] and research orientated - digitized - workflows [5].

For scientists, administering a local RIS always comes with extra effort. At FAU, we try to offer several advantages: For example, FAU CRIS acts as content source for web pages (publications, projects, infrastructure, navigation structure based on research fields, ...) simplifying maintenance for personal, chair, department and faculty presentations.

Adding new entity types to Converis doesn't require software development, a user-defined data model is a key feature. So, this fundamental step is not an issue, it turned out that most challenges can be found in meta data complexity and workflow definition, including user interaction.

2 Motivation and challenges

It's very easy to publish a research dataset in a global repository, like Zenodo [6]: Upload data, fill in some attributes and you're done. On the other hand, at this stage the institutional RIS is not aware of this new information, so it cannot be included in project reports, personal web pages, etc. automatically. New promising approaches like the PID Graph [7] try to tackle those issues, but they need a global PID coverage, which is evolving, but is currently incomplete.

Furthermore, in the research data life cycle, sometimes it's not permitted or reasonable to publish data or even store it outside the own institution. But from funders' perspective, professional data management is mandatory, so that universities need to offer an institutional data repository ensuring data availability. Since research data should be FAIR [8] in the end, data producers need to supply rich meta data.

If data is to be findable and accessible, it has to be equipped with meta data. Additionally, for the data to satisfy interoperability, the meta data has to follow a meta data schema, which is ideally commonly agreed upon. To gain a maximum of flexibility on further data usage and publication, we target the

DataCite schema internally from the very beginning since it is used more broadly as other schemas, e.g. the Data extension in CERIF [9]. Actually, we don't implement the schema as a whole in FAU CRIS, we're able to generate it out of CRIS' data. Currently, we are using the most recent DataCite Metadata Schema V4.4 (released 30 Mar 2021) [10].

This schema separates properties into three different levels of obligation: mandatory, recommended and optional. Since we judged some recommended properties to be of such importance, we decided to make them mandatory when entering meta data in CRIS. This includes, for example, the description of the data, which is, in our opinion, essential to any future re-use. Some properties offer great value in exchange for a small amount of user interaction, like for example the image type, therefore they were made mandatory, too.

In exchange for forcing a user to enter recommended properties towards richer meta data, we automatically generate several properties, thereby eliminating as much as possible of the manual entry of meta data. This includes basic file information like the file size, checksum or MIME type, but also, by using CRIS, we have access to a lot of information like e.g., the ROR of FAU or the authors' IDs.

3 Solution

In order to support the FAIR principles for datasets and their embedment in FAU' research information system as much as possible, we set up a combination of a software tool that is run on the researchers' local computer and FAU CRIS.

The process of archiving data is broken into four steps:

1. Define dataset
2. Generate file metadata
3. Upload meta data into FAU CRIS and enrich it towards DataCite schema
4. Upload to institutional repository (FAUDataCloud)

Everything except step no. 3 is covered by FAU's *Lightweight datacloud upload tool (FAU_L)*. So, *FAU_L* supports our users on dataset preparation, data integrity and easy cloud storage access.

The application is a multi-platform Python program using a combination of *Eel* [11] and *VUE* [12] to offer a user-friendly and barrier-free interface. The GUI supplies the user with tooltips upon hovering elements, supports multiple languages and warns the user about mistakes or slips (Fig 1).

If an unnoticed error occurs during the archiving process of a dataset, it may no longer satisfy the FAIR principles. Using checksums, we can verify, that an uploaded file arrived correctly.

As the first step, a user has to decide about which files go into a dataset. Second, they use *FAU_L* to automatically generate a relevant file metadata, like name, size and mime-type. This metadata is used to pre-fill as much as possible of the DataCite schema in FAU CRIS.

For the third step, the user accesses FAU CRIS, uploads the meta data file generated by *FAU_L* and fills out the remaining mandatory properties in FAU CRIS (like description, authors, etc.) Only after that, the user will receive an upload token to be entered in *FAU_L*. Using the open source, secure and encrypted web application *FileSender* [13] the dataset is uploaded into a staging area as the fourth and final step. Once the upload is verified, it's moved to our FAUDataCloud institutional storage.

After upload, the user decides on visibility and accessibility of the dataset directly in FAU CRIS:

1. not visible at all to the public (neither meta data nor the data itself)
2. only meta data visible

3. open data

Once meta data was made available to the public, a DOI is assigned automatically. This implies that this decision cannot be undone. FAUDataCloud allows data sharing in all stages using private links independently of FAU CRIS' visibility setting.

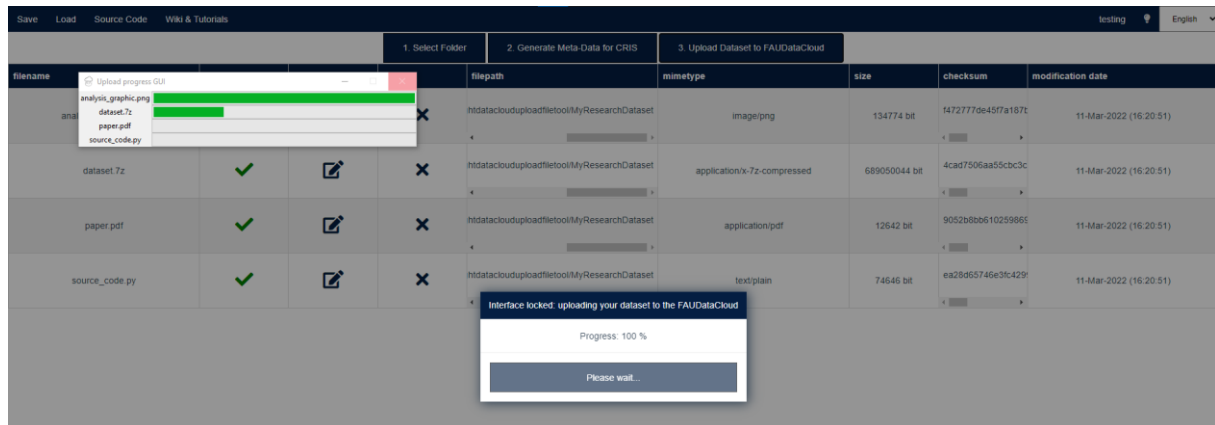


Fig 1: *FAU_L* uploading a dataset to the staging area. We can see the progress bars for individual files and their metadata in the slightly grayed area of the locked interface.

4 Summary and Outlook

We set up a workflow for handling FAIR research data, including their significance for research information. This includes crosslinking to all other entity types usually present in research information systems.

Furthermore, using the established rights and roles definitions, responsibilities and management in FAU CRIS (including chair admins, delegations, organizational structures), we ensure that even unpublished datasets will be accessible in case all authors have left the university.

FAU_L is currently in pilot stage. Roll-out for whole FAU is scheduled for late Q2/2022. Heavy users will be addressed by a command line interface.

The availability of dataset's overarching meta information allows us at a later stage the relocation of published datasets to global or discipline specific repositories in order to enhance the dataset's FAIRness.

References

- [1] Melsheimer, B., & Walther, M. (2016). Introducing CRIS at FAU. *Procedia Computer Science*, 106, 239-244. <https://dx.doi.org/10.1016/j.procs.2017.03.021>
- [2] Asserson, A., Jeffery, K. G., & Lopatenko, A. (2002). CERIF: Past, Present and Future: An Overview. In „Gaining Insight from Research Information“: Proceedings of the 6th International Conference on Current Research Information Systems (S. 33–40). euroCRIS; Kassel University Press.
- [3] <https://www.kerndatensatz-forschung.de/> (fetched on Mar 13th 2022)
- [4] Walther, M., & Melsheimer, B. (2019). Automated author affiliation processing using Scopus data. *Procedia Computer Science*, 146, 53-59. <https://dx.doi.org/10.1016/j.procs.2019.01.079>
- [5] <https://www.fau.eu/?p=1316165> (fetched on Mar 13th 2022)

- [6] <https://zenodo.org/> (fetched on Mar 13th 2022)
- [7] Fenner, M., & Aryani, A. (2019). Introducing the PID Graph (Version 1.0). <https://doi.org/10.5438/JWVF-8A66>
- [8] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016). <https://doi.org/10.1038/sdata.2016.18>
- [9] https://eurocris.org/eurocris_archive/cerifsupport.org/ (fetched on Mar 13th 2022)
- [10] https://schema.datacite.org/meta/kernel-4.4/doc/DataCite-MetadataKernel_v4.4.pdf (fetched on Mar 13th 2022)
- [11] <https://github.com/ChrisKnott/Eel> (fetched on Mar 13th 2022)
- [12] <https://vuejs.org/> (fetched on Mar 13th 2022)
- [13] <https://filesender.org/> (fetched on Mar 13th 2022)