A FAIR Digital Object Lab Software Stack

Andreas Pfeil‡, Thomas Jejkal‡, Sabrine Chelbi‡, Nicolas Blumenröhr‡

‡ Karlsruhe Institute of Technology, Karlsruhe, Germany

Corresponding author: Andreas Pfeil (andreas.pfeil@kit.edu)

Abstract

Preprocessing data for research, like finding, accessing, unifying or converting, takes up to large parts of research time spans (Wittenburg and Strawn 2018). The FAIR (Findability, Accessibility, Interoperability, Reusability) principles (Wilkinson 2016) aim to support and facilitate the (re)use of data, and will contribute to alleviating this problem. A FAIR Digital Object (FAIR DO) captures research data resources of all kinds (raw data, metadata, software, ...) in order to align them with the FAIR principles.

FAIR Digital Objects are expressive, machine-actionable pointers to research data (De Smedt et al. 2020). As such, each FAIR DO points to one research data object. Additionally, they may link to other FAIR DOs, explaining their relations. The FAIR Digital Object Lab (Pfeil et al. 2022) is an extendable and adjustable architecture (a software stack) for generic FAIR Digital Object tasks. It consists of a set of interacting components with services and tools for creation, validation, discovery, curation, and more. In this talk, we will present our plans for the FAIR DO Lab and explain our decisions, which are mostly based on the experience gained in previous developments.

The creation and maintenance of FAIR DOs is not trivial, as their persistent identifiers (PIDs) contain typed record information. When creating or maintaining PID records of FAIR DOs, the required information has to be validated, involving calls to a public Data Type Registry (DTR) (Lannom et al. 2015). After a successful validation, the information has to be transformed to a representation of a PID service. After a FAIR DO has been registered successfully, the PID should be documented locally and disseminated. Using these PIDs as a starting point, tools may use the machine-actionability of FAIR DOs to maintain search indexes or to create collections. This enables researchers to look up PIDs by searching for record information or timestamp.

We are developing a set of services, offering a solution to support these use-cases, which we call the FAIR DO Lab. Its goal is to have a production-ready and configurable software stack, easing the development of FAIR-DO-aware tools and services by offering at least the described use-cases. We have already gained some experience by its predecessor, the FAIR DO Testbed (Pfeil et al. (2021a)), which was introduced at the Research Data Alliance (RDA) Virtual Plenary 17 Poster Session (Pfeil et al. 2021b). The Lab will be

configurable similar to the Testbed, as each service can be omitted or replaced to satisfy specific needs while integrating the Lab on top of existing research infrastructures.

The FAIR DO Lab enables PID record management and validation using the Typed PID Maker (Pfeil and Jejkal 2021), following the RDA PID Information Types (PIT) Working Group Recommendations (Weigel et al. 2015) and an external Data Type Registry (DTR), following the RDA Data Type Registry Working Group Recommendations (Lannom et al. 2015). The DTR stores profiles and types, enabling typed, machine-actionable PID records. The Typed PID Maker uses this information for the validation of PID records, and stores and disseminates PIDs after their creation.

All created or modified PIDs are communicated to a message broker. This way, other services can be notified about such activities. Our first service making use of this will be an advanced indexing service. It will ingest the PIDs and their record information into a search index, but also try to extract information from the bit-sequence of the digital object itself. In a second step, we are considering the automated creation of collections utilizing our production-ready Collection Registry (Chelbi and Jejkal 2020), which the Testbed already includes. This will require a set of rules and a process to use those rules in order to place new PIDs in the correct collection. The Collection Registry is an implementation of the Collection API specification (Weigel et al. 2017), which was published by the corresponding RDA Research Data Collections Working Group.

On the conceptual side, we hope to gain more insight about the required structure of PID records. There are ongoing discussions about this structure and to which degree standardization is required. Large talking points are the concepts of Digital Object Types (Lannom et al. 2015) and Kernel Information Profiles (Weigel et al. 2018). Working on the Lab and its predecessor, we recognized that there are large gaps regarding the structure of FAIR Digital Objects and the roles of the object types and profiles. To bring FAIR DOs into reality, research software will need to use them. But as FAIR DOs point to diverse kinds of research data, the software needs to make decisions. To what extent can the software use a specific FAIR DO? We observed that too much flexibility makes automated decisions harder. Our suggestion is therefore to consider FAIR DOs less from the infrastructure point of view, and more from the machine's point of view to improve the machine-actionability. We expect that we will gain insights about the feasibility in the development process to ease the development of further FAIR-aware tools for research, particularly for specialized tools that already exist and are in use. It will not be feasible to write every tool from scratch.

On the practical side, the Lab will already have a stronger focus on interactive tools with user interfaces in order to provide an easy-to-use Lab for research. We consider our current work on granular base services for research data management to be a solid ground for such developments. These tools can of course not replace specialized tools, but will make the generic services in the Lab easy to use. We still expect that specialized tools will benefit from the integration of such services.

The FAIR DO Lab development has been supported by the research program ‘Engineering Digital Futures’ of the Helmholtz Association of German Research Centers and the Helmholtz Metadata Collaboration Platform.

Keywords

FAIR Digital Objects, Tools, Persistent Identifiers, Kernel Information Profiles, Research Environments, Helmholtz Metadata Collaboration Platform (HMC)

Presenting author

Andreas Pfeil

Presented at

First International Conference on FAIR Digital Objects, presentation

Conflicts of interest

References

- Chelbi S, Jejkal T (2020) Collection Registry. URL: https://github.com/kit-data-manager/collection-api
- De Smedt K, Koureas D, Wittenburg P (2020) FAIR Digital Objects for Science: From Data Pieces to Actionable Knowledge Units. Publications 8 (2). https://doi.org/10.3390/publications8020021
- Lannom L, Broeder D, Manepalli G (2015) RDA Data Type Registries Working Group Output. Zenodo. https://doi.org/10.15497/A5BCD108-ECC4-41BE-91A7-20112FF77458
- Pfeil A, Jejkal T, et al. (2021) Typed PID Maker. URL: https://github.com/kit-data-manager/pit-service
- Pfeil A, Jejkal T, Chelbi S, et al. (2021a) FAIR Digital Object Testbed. URL: https://github.com/kit-data-manager/testbed4inf
- Pfeil A, Jejkal T, Chelbi S, Stotzka R (2021b) FAIR Digital Object Ecosystem Testbed. Karlsruher Institut für Technologie (KIT). https://doi.org/10.5445/IR/1000131613
- Pfeil A, Jejkal T, Chelbi S, et al. (2022) FAIR Digital Object Lab. URL: https://github.com/kit-data-manager/FAIR-DO-Lab
- Weigel T, DiLauro T, Zastrow T (2015) PID Information Types WG final deliverable. Research Data Alliance https://doi.org/10.15497/fdaa09d5-5ed0-403d-b97a-2675e1ebe786
- Weigel T, Almas B, Baumgardt F, Zastrow T, Schwardmann U, Hellström M, Quinteros J, Fleischer D (2017) RDA Research Data Collections WG Recommendations. Research Data Alliance https://doi.org/10.15497/rda00022

- Weigel T, Plale B, Parsons M, Zhou G, Luo Y, Schwardmann U, Quick R, Hellström M, Kurakawa K (2018) Recommendation on PID Kernel Information (Version 1). Research Data Alliance. https://doi.org/10.15497/rda00031
- Wilkinson M, et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3 (1). https://doi.org/10.1038/sdata.2016.18
- Wittenburg P, Strawn G (2018) Common Patterns in Revolutionary Infrastructures and Data. https://doi.org/10.23728/b2share.4e8ac36c0dd343da81fd9e83e72805a0