Realizing FAIR Digital Objects for the German Helmholtz Association of Research Centres

Thomas Jejkal[‡], Andreas Pfeil[‡], Jan Schweikert[‡], Anton Pirogov[§], Pedro Videgain Barranco[§], Florian Krebs[|], Christian Koch[¶], Gerrit Guenther[#], Constanze Curdt[¤], Martin Weinelt[¤]

‡ Karlsruhe Institute of Technology, Eggenstein-Leopoldshafen, Germany
§ Forschungszentrum Juelich, Juelich, Germany
| German Aerospace Center, Bonn, Germany
¶ German Cancer Research Center, Heidelberg, Germany
# Helmholtz-Zentrum Berlin, Berlin, Germany
¤ Geomar, Kiel, Germany

Corresponding author: Thomas Jejkal (thomas.jejkal@kit.edu)

Abstract

The Helmholtz Association (Anonymous 2022d), the largest association of large-scale research centres in Germany, covers a wide range of research fields employing more than 43.000 researchers. In 2019, the Helmholtz Metadata Collaboration (HMC) (Anonymous 2022f) Platform as a joint endeavor across all research areas of the Helmholtz Association was started to make the depth and breadth of research data produced by Helmholtz Centres findable, accessible, interoperable, and reusable (FAIR) for the whole science community. To reach this goal, the concept of FAIR Digital Objects (FAIR DOs) has been chosen as top-level commonality for existing and future infrastructures of all research fields.

In doing so, HMC follows the original approach of realizing FAIR DOs based on globally unique, Persistent Identifiers (PID), e.g., provided by https://handle.net/, machine actionable PID Records and strong typing using Data Types like https://dtr-test.pidconsortium.eu/#objects/21.T11148/1c699a5d1b4ad3ba4956 registered in a Data Type Registry, e.g., http://dtr-test.pidconsortium.eu/. In all these areas, HMC can build on the great groundwork of the Research Data Alliance and the FAIR DO Forum. However, when it comes to realization, there are still some gaps that will have to be addressed during our work and will be raised in this presentation.

For single FAIR DO components like PIDs and Data Types, existing infrastructures are already available. Here, the Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG) (Anonymous 2022e) provides strong support with their many years of experience in this field. Within the framework of the ePIC consortium (Anonymous 2022c), the GWDG is offering on the one hand PID prefixes based on a sustainable business model, on the other hand GWDG is very active in terms of providing base services required for realizing FAIR DOs, e.g., different instances of Data Type Registries for accessing,

creating, and managing Data Types required by FAIR DOs. Besides that, in the context of HMC we developed a couple of technical components to support the creation and management of FAIR DOs: The Typed PID Maker (Pfeil 2022b) providing machine actionable interfaces for creating, validating, and managing PIDs with machine-actionable metadata stored in their PID record, or the FAIR DO testbed, currently evolving into the FAIR DO Lab (Pfeil 2022a), serving as reference implementation for setting up a FAIR DO ecosystem. However, introducing FAIR DOs is not only about providing technical services, but also requires the definition and agreement on interfaces, policies, and processes.

A first step in this direction was made in the context of HMC by agreeing on a Helmholtz Kernel Information Profile (http://dtr-test.pidconsortium.eu/#objects/21.T11148/b9b76f887845e32d29f7). In the concept of FAIR DOs, PID Kernel Information as defined by Weigel et al. (Weigel et al. 2018) is key to machine actionability of digital content. Strongly relying on Data Types and stored in the PID record directly at the PID resolution service, PID Kernel Information can be used by machines for fast decision making. The Helmholtz Kernel Information Profile is an attempt to introduce a top-level commonality across all digital assets produced within the Helmholtz Association and beyond to establish a basis for FAIR research data based on FAIR DOs.

Hereby, the Helmholtz Kernel Information Profile integrates the recommendations of the RDA PID Kernel Information Working Group (Anonymous 2022b) as far as possible. By extending the Draft Kernel Information Profile (Weigel et al. 2018) with additional, mostly optional attributes, the Helmholtz Kernel Information Profile allows the adding of contextual information to FAIR DOs, e.g., research topic, or contact information, which is then available for machine decisions. Furthermore, additional properties for representing relationships between FAIR DOs, e.g, hasMetadata and isMetadataFor, were introduced to allow mutual relations between FAIR DOs.

Currently, a demonstrator is implemented integrating the above components and services, i.e., PID Service, Data Type Registry, and Typed PID Maker. Fig. 1 outlines the architecture overview of the first version of the demonstrator.

In this first version, in a semi-automatic workflow, a user enters a Zenodo (Anonymous 2022a) PID in a graphical Web frontend. A mapping component tries to fill automatically at least the properties required by the Helmholtz Kernel Information Profile using the obtained Zenodo metadata record. In a manual validation loop, the user may add or update certain properties before they are sent to an instance of the Typed PID Maker, validated against the Helmholtz Kernel Information Profile, and stored in the record of a newly registered PID using the services of the ePIC consortium. In addition, registered PID records are made searchable via the graphical frontend on top of a search index, e.g., realized using https://www.elastic.co/.

After implementing this generic workflow, additional mappers supporting other repository platforms will be implemented based on the lessons learned, which will lead to a growing number of FAIR DOs and holds potential for providing significant benefits to scientists, e.g., a central point of contact for research data sets stored in different repositories, machine-

actionable identification of relevant datasets, and creation of knowledge graphs representing relationships between data sets, repository platforms, researchers and research organizations.

Furthermore, the gathered experience and its documentation will help others to apply the FAIR DO concept more easily, which will lead to an ever-growing collection of available FAIR DOs with an increasing quality and level of automation at creation time.

Keywords

Helmholtz Metadata Collaboration Platform, Persistent Identifiers, PID Kernel Information Profile, Demonstrator

Presenting author

Thomas Jejkal

Presented at

First International Conference on FAIR Digital Objects, presentation

Conflicts of interest

References

- Anonymous (2022a) Zenodo. URL: https://zenodo.org/
- Anonymous (2022b) PID Kernel Information Working Group. URL: https://www.rd-alliance.org/groups/pid-kernel-information-wg
- Anonymous (2022c) Persistent Identifiers for eResearch (ePIC). URL: https://www.pidconsortium.net/
- Anonymous (2022d) Helmholtz Association. URL: https://www.helmholtz.de/
- Anonymous (2022e) Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen. URL: https://www.gwdg.de/
- Anonymous (2022f) Helmholtz Metadata Collaboration Platform. URL: https://helmholtz-metadaten.de/en
- Pfeil A (2022a) FAIR DO Lab. URL: https://github.com/kit-data-manager/testbed4inf
- Pfeil A (2022b) Typed PID Maker. URL: https://github.com/kit-data-manager/pit-service
- Weigel T, Plale B, Parsons M, Zhou G, Luo Y, Schwardmann U, Quick R, Hellström M, Kurakawa K (2018) RDA Recommendation on PID Kernel Information. Online https://doi.org/10.15497/RDA00031

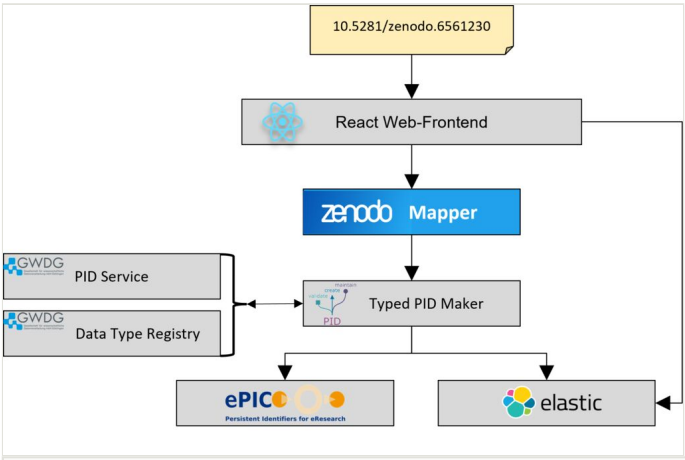

Figure 1.

Architecture of the FAIR DO demonstrator.