EPiC Series in Computing

Volume 105, 2025, Pages 284–292

Proceedings of EUNIS 2024 annual congress in Athens

FAIR Digital Objects: FAIRtilizer for the Digital Harvest

Marius Politze[1*], Benedikt Heinrichs[1†], Sirieam Hunke[1‡], Ilona Lang[1§] and Thomas Eifert[1**]

[1] RWTH Aachen University, Germany

politze@itc.rwth-aachen.de, heinrichs@itc.rwth-aachen.de, hunke@itc.rwth-aachen.de, lang@itc.rwth-aachen.de, eifert@itc.rwth-aachen.de

Abstract

FAIR Digital Objects (FDO) are a concept to transfer digital objects to the world of the FAIR principles (Findable, Accessible, Interoperable, Re-Usable) that form a common baseline for the handling of research data. Combined with an enterprise-ready cloud storage system, FDOs can be used to make these systems fit for purpose in the research data management (RDM) context. This allows profiting from scalability by connecting data spaces concepts defined in Gaia-x. The presented concepts are implemented based on a shared, geo-redundant storage system and within the research data management platform Coscine that is made available to researchers in the German federal state of North Rhine-Westphalia.

1 Preparation of the Soil

National Initiatives – like the German NFDI consortia are devoted to setting discipline-specific standards and requirements for research data management (RDM). They define on a national level what standards should be adhered to when managing research data. However, it remains up to the research institutions to provide the technical means for the researchers to implement these standards.

For the federal state of North-Rhine-Westphalia (NRW), a consortium consisting of the universities RWTH Aachen University, TU Dortmund, University of Duisburg-Essen, University of Cologne, and Ruhr University Bochum implemented a shared research data storage infrastructure that can now be used by these universities themselves, 16 universities of applied sciences and the 7 universities of arts

[*] https://orcid.org/0000-0003-3175-0659
[†] https://orcid.org/0000-0003-3309-5985
[‡] https://orcid.org/0000-0001-9316-4220
[§] https://orcid.org/0000-0002-7202-5982
[**] https://orcid.org/0000-0003-1996-0944

and music in NRW and by their collaboration partners. Using an enterprise grade general-purpose storage system, the consortium thus laid a first base for the implementation of various best practices defined by the RDM communities.

Within the “Research Data Storage for NRW“ (RDS.NRW) each consortium member operates a share of a geo-distributed system that is logically separated into geographical access zones (see Figure 1):

- local at Aachen (with three subsidiaries)
- local at Duisburg-Essen
- local at Dortmund (with two subsidiaries)
- local at Cologne (with three subsidiaries)
- local at Bochum
- geo-redundant at Aachen, Duisburg-Essen and Cologne

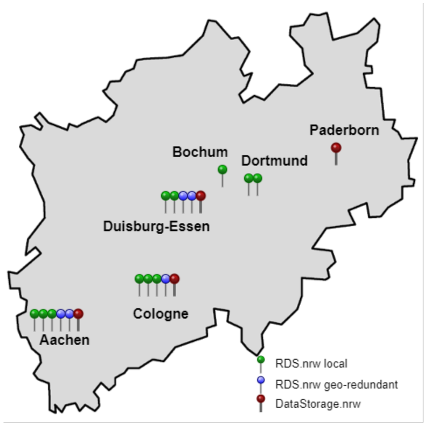

Figure 1: Storage federation across the state of NRW.

While these zones are technologically identical, researchers could choose based on proximity and/or requirements for redundancy.

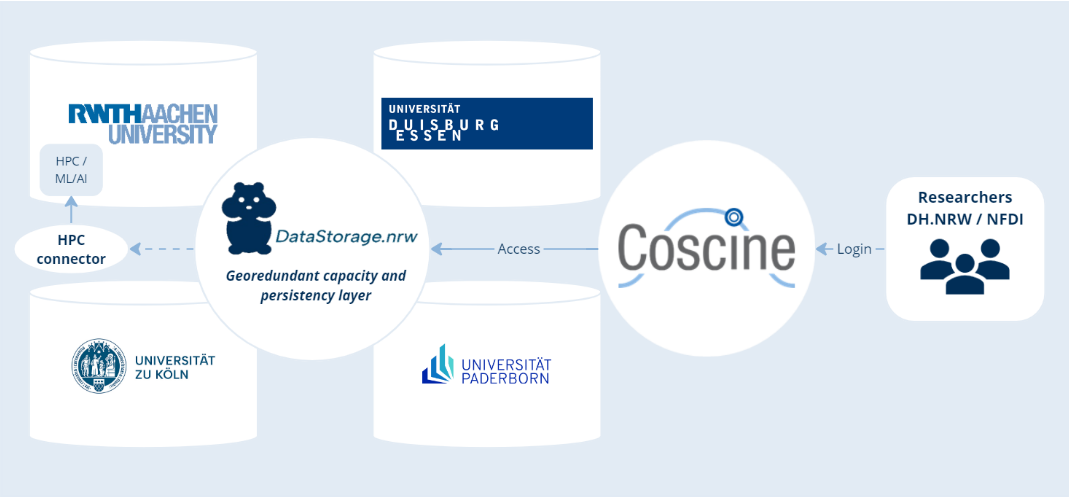

Figure 2: Layout of the geo-replicated DataStorage.nrw with local connector for HPC and AI access.

This first system was a great success, and it paved the way to overcome organizational aspects when offering a geo-redundant storage system by multiple universities. It is therefore a blueprint for the next generation: "DataStorage.nrw" that will raise the overall quality of the storage service by exclusively offering all geo-redundant storage located at RWTH Aachen University, University of Duisburg-Essen, University of Cologne, and University of Paderborn and a dedicated high-performance tier – so called HPC connector for large scale computing and AI access for the national compute cluster at Aachen (see Figure 2).

Both systems, RDS.NRW and DataStorage.nrw are extremely sophisticated – industry grade – infrastructures and offer very high data security, performance, and a good capacity to cost ratio. However, the systems do not consider integration into the scientific world: A surrounding set of management processes needed to be put into place for distribution and allocation as well as for the FAIR data management on the systems.

2 Planting the Seeds

The storage space is distributed using a science led process that is heavily inspired by the distribution of computing time in (national) high performance computing centers. Researchers from all participating universities apply equally and self-determined for storage quota and resources are granted based on scientific impact and structural fitness, e.g., based on a data management plan. The provisioning process is performed using JARDS as an application management system (Lang, Nellesen, Bossert, & Politze, 2023) and the data management platform Coscine is used for allocation of the resources (Politze, et al., 2020). Applications for storage grants are centrally peer-reviewed by a pool of researchers and data managers originating from all participating universities. This "science-led" distribution process ensures that storage space is distributed transparently, and processes are equally accessible for researchers from all institutions, even if they are not participating in the consortium operating the storage systems. Additionally, the process ensures that all provided resources are associated with a respective scientific purpose.

Secondly, the data stored in the system needs to be managed according to the FAIR principles (Wilkinson, et al., 2016). In short, the FAIR principles refer to:

- Findable: assigning a persistent identifier (PID) and indexing of metadata
- Accessible: allowing transfer of data using standardized communications protocols
- Interoperable: usage of applicable formats and vocabularies for (meta)data representation
- Re-Usable: assignment of provenance, terms, and licenses

This is supported by Coscine: in the allocation process each project and resource is assigned with a set of metadata and a PID in the Handle system (Kálmán, Kurzawe, & Schwardmann, 2012) that is globally unique and resolvable. Researchers can individually manage access rights through the AAI federation eduGAIN. Files within the resource may be described with more sophisticated metadata according to discipline-specific metadata profiles (Grönewald, et al., 2022). All information is validated and stored in a linked data knowledge graph within Coscine and can be managed by the researchers through a web interface or a set of REST APIs. The data itself can be accessed directly on the storage system to retain the best possible performance.

In this sense, Coscine governs a structured hoard of research data and metadata. In contrast to more traditional data repositories, however, data can be changed at any time by the researchers (more like a git repository) and the metadata profiles allow discipline-specific description and validation of stored metadata.

3 Supporting the Growth

Simply having Coscine govern the data is not sufficient to serve the vision of a truly connected research data ecosystem. Several other choices of software co-exist, some of which are much more discipline-specific like electronic lab notebooks, e.g., eLabFTW, or much more generic like repositories, e.g., Zenodo. While it is certainly possible and intended to hook up other storage systems to Coscine, its metadata database is intended to feel like a single point of access, in the sense of a community cloud service. This operating model might not fit for projects with very high confidentiality. For these various reasons, it is a clear aim within the NFDI consortia to assure an interoperability layer between different data and metadata management systems. For Coscine and two associated discipline-specific NFDI consortia NFDI4Ing (engineering sciences) and NFDI-MatWerk (material sciences) the FAIR Digital Object (FDO) concept was chosen as an initial model. One of the reasons for choosing FDO was that it is defined as a requirement for "turning FAIR into reality" by the Directorate General for Research and Innovation of the European Commission (Directorate General for Research and Innovation, 2018). Furthermore, the Research Data Alliance (RDA) endorses an FDO based implementation in the Handle PID system (Weigel, et al., 2018).

In general, an FDO is defined as a machine-readable and machine actionable unit of data that is identified by a PID and described by a record. The FDO record therefore should include the most critical information about the objects' context and possible operations. The context should be a minimal set of metadata, in the sense that it should allow discovery and re-usability. Operations define how to access the binary data and further metadata – they can be as simple as a single HTTP GET (considering a single file) or a description of a complex interface (e.g., for a Web Map Tile Service or a SPARQL endpoint). In this sense, a FDO is a technical incarnation of the FAIR principles (see Figure 3):

- The PID makes the record findable.
- Operations based on general protocols make data and metadata accessible.
- A broadly understandable metadata record and context in the record makes it interoperable.
- The record indicates re-usability of the data.

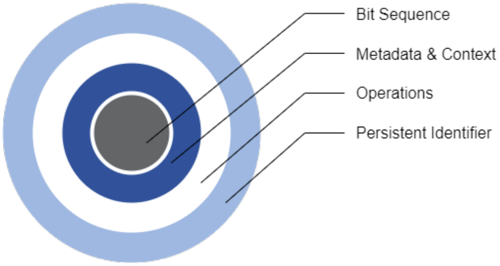

Figure 3: A FAIR Digital Object is a unit of information from a bit sequence, metadata, operations identified by a persistent identifier (PID).

3.1 Cultivation

The technical architecture of Coscine defines layers of abstraction to translate the basic interfaces of the storage system to the more high-level interfaces as defined by the FDO. This is especially not limited to accessing the binary data or the metadata of the digital object but also includes means for authorization. The underlying pSTAIX architecture (Politze, Decker, & Eifert, 2017) serves as a model for the implementation (see Figure 4):

- Tier 0 contains authentication and authorization based on the identity federation of eduGAIN (López, 2006) and other "social sign on" providers. Most importantly, ORCID (Haak, Fenner, Paglione, Pentz, & Ratner, 2012) as a PID for researchers across their life cycle (e.g., when changing institutions).

- Tier 1 contains temporary local storage for session states, intermediate results, and logging not directly visible to researchers using Coscine.
- Tier 2 contains the interfaces of the attached storage systems – most prominently RDS.NRW and DataStorage.nrw. Both systems implement a subset of the S3 command set known from AWS's cloud offer. This could be extended to other storage appliances and interfaces in case of migration or extension to other (geographic) communities. The tier also includes the internal graph database that Coscine accesses via a SPARQL interface.
- Tier 3 are internal abstractions for the development of Coscine that bundle operations overarching the storage systems and database, ensuring authentication, and validating information provided by the users.
- Tier 4 is the generic REST API provided by Coscine that is used by the web user interface and applications or scripts implemented by scientists for automation. These APIs can also be used to integrate Coscine into other scientific working environments, like electronic lab notebooks.
- Tier 5 abstracts the individual operations even further by a clear process-oriented definition. As such, the interfaces in this tier focus on PIDs, records, context, and most importantly on translating the operations to interfaces of tier 4 and below.

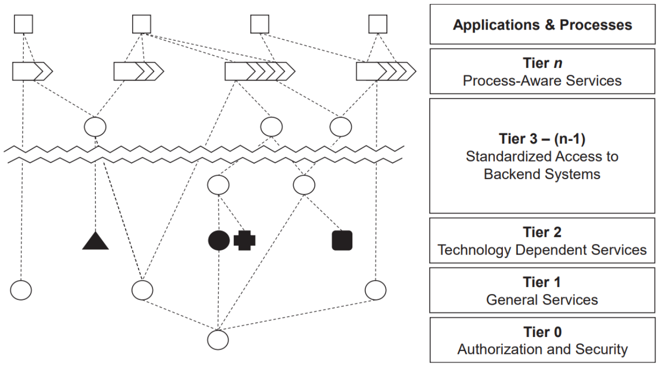

Figure 4: The pSTAIX reference architecture (Politze, Decker, & Eifert, 2017).

Hence, tier 5 defines the abstraction layer for the implementation of the FDO concept within Coscine.

3.2 FAIRtilizing Research Data Storage with Coscine

As previously mentioned, the FDO defines a concept that it is widely technology independent. Consequently, there are several technology stacks available that meet the requirements for the implementation of the FDO concept – each with a slightly different focus. Within Coscine we decided to implement the FDO concept using two mostly independent approaches: First, the Kernel Information Record (KIR) based on the Handle PID system as recommended by the RDA (Weigel, et al., 2018) and second the FAIR Data Point (FDP) (Bonino, Burger, & Kaliyaperumal, 2022) based on the W3C standards Linked Data Platform (LDP) (Speicher, Arwe, & Malhotra, 2015) and the Data Catalog Vocabulary (DCAT) (Maali & Erickson, 2014) that offers compatibility to the International Data Space (International Data Spaces Association, 2023) and Gaia-x. These two quite different approaches were chosen to demonstrate their compatibility and most importantly their interoperability in the context of FAIR data.

The KIR based implementation stores the FDOs record as typed properties directly within the PID information. This automatically replicates all information within the Handle network and makes it accessible directly from the resolver. This minimizes the required number of HTTP requests and the load on the data platform. The different types of properties as well as a Kernel Information Profile (KIP) or type of the FDO are registered in a central type registry. The registry allows the discovery of specifications of properties as well as operations associated with the KIP of the FDO at hand. Storage resources in Coscine use a tailored KIP containing minimal information about owner, usage license and links to APIs and storage system endpoints to retrieve metadata in RDF format (Cyganiak, Wood, & Lanthaler, 2014) and list the root directory of the storage resource.

Secondly, the FDP based implementation uses the metadata in RDF format as records. To separate the entry points and not cause confusion on the clients' end, PURLs are used as PIDs in this case. Resolving the PURL will result in an HTTP redirect to the metadata document. This is slightly less economical than the previous approach as this technically results in two requests and load on both the resolver, and the data platform. However, the returned metadata document is much more extensive, as DCAT provides links to discover individual files and their metadata within the resource. The underlying LDP implementation defines all operations needed to automatically discover the contents of the resource.

Interoperability is achieved by defining LDP based operations in the KIP and therefore aligning both approaches with minimal effort on the interfaces.

4 Harvesting Fruits of Success

Obviously, the FDO is motivated from a RDM perspective, with the long tail and data availability in mind. How adhering to the FAIR principles is beneficial for scientific output is being widely discussed in various scientific disciplines and is certainly important for the adoption of RDM tools like Coscine. However, looking back at the initially introduced storage system, how can an IT systems' provider profit from implementing the FAIR principles with FDO.

4.1 Diversification of Crops

The FDO concept defines a very high-level and abstract interface to access the data. Additionally, both approaches operate merely on links that can (or must) be resolved using only a few entry points defined by the PIDs. This makes it especially easy to hide a storage systems' complexity for the researchers. While Coscine allows complex management processes for provisioning, allocation, access management and metadata validation, none of these interfere with the FDO implementation. On the contrary, all (read) operations could be implemented using a simple web server and static binary files for bit sequences and RDF files for metadata.

This can greatly increase the diversification of specialized and decentralized storage services operated for or by individual research groups. Even further, a precocious agreement on an FDO implementation can allow a fluent migration of data from decentralized to centralized storage services according to a data life cycle or if requirements or funding schemes require doing so. This also helps automated processes e.g., for AI training or other data science techniques to automate data acquisition.

Lastly, it opens the door for future more loosely coupled storage federations as technological dependencies are widely abstracted (Politze, et al., 2023).

4.2 Preparation for the Next Season

Transitioning between technologies is not only limited to cooperations with other institutions. Considering current retention times of at least ten years for raw research data and ever-increasing storage capacities and data volumes, migrations will become significantly more time-consuming. We see that abstraction also alleviates migrations from one generation of a storage system to the next. Reducing the dependencies on the specific technology dependent interfaces for data access, both internally and for researchers using the system, allows better incremental migration scenarios.

While the standards defined by the W3C and picked up in the protocols of International Data Space and Gaia-x have the potential to gain significant traction in industrial applications over the next few years, the FDO concept remains technology independent and should outlive current implementations.

5 Acknowledgements

RDS and RDS.NRW are funded by Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen (MKW:124-4.06.05.08-139057, DFG: INST222/1261-1). DataStorage.nrw is funded by Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen (MKW: 214-76.01.09-7-7937 DFG: INST 222/1530-1).

The conceptual work was supported with resources granted by NFDI4Ing, funded by Deutsche Forschungsgemeinschaft (DFG) under project number 442146713, NFDI-MatWerk, funded by Deutsche Forschungsgemeinschaft (DFG) under project number 460247524 and FAIR Data Spaces, funded by the German Federal Ministry of Education and Research (BMBF) under funding reference FAIRDS11.

6 References

Bonino, L., Burger, K., & Kaliyaperumal, R. (Eds.). (2022, August 26). FAIR Data Point. *FAIR Data Point*. Retrieved January 25, 2023, from https://specs.fairdatapoint.org/

Cyganiak, R., Wood, D., & Lanthaler, M. (Eds.). (2014). RDF 1.1 Concepts and Abstract Syntax. *RDF 1.1 Concepts and Abstract Syntax*. Retrieved February 20, 2020

Directorate General for Research and Innovation. (2018). *Turning FAIR into reality: final report and action plan from the European Commission expert group on FAIR data.* European Commission. doi:10.2777/1524

Grönewald, M., Mund, P., Bodenbrenner, M., Fuhrmans, M., Heinrichs, B., Müller, M. S., . . . Stäcker, T. (2022). Mit AIMS zu einem Metadatenmanagement 4.0: FAIRe Forschungsdaten benötigen interoperable Metadaten. In V. Heuveline, & N. Bisheh (Eds.), *E-Science-Tage 2021.* Heidelberg, Germany: heiBOOKS. doi:10.11588/heibooks.979.c13721

Haak, L. L., Fenner, M., Paglione, L., Pentz, E., & Ratner, H. (2012). ORCID: a system to uniquely identify researchers. *Learned Publishing, 25*, 259–264. doi:10.1087/20120404

International Data Spaces Association. (2023). Dataspace Protocol - Working Draft. *Dataspace Protocol - Working Draft*. Retrieved February 8, 2024, from https://docs.internationaldataspaces.org/ids-knowledgebase/v/dataspace-protocol

Kálmán, T., Kurzawe, D., & Schwardmann, U. (2012). European Persistent Identifier Consortium - PIDs für die Wissenschaft. In R. Altenhöner, & C. Oellers (Eds.), *Langzeitarchivierung von Forschungsdaten* (pp. 151–164). Berlin, Germany: Scivero Verl.

Lang, I., Nellesen, M., Bossert, L. C., & Politze, M. (2023). Carrots and Sticks: Motivating with Storage for Good RDM – Science Led Allocation of Research Data Storage Resources within an Integrated RDM System. In V. Heuveline, N. Bisheh, & P. Kling (Eds.), *E-Science-Tage 2023: Empower Your Research – Preserve Your Data.* heiBOOKS. doi:10.11588/HEIBOOKS.1288.C18071

López, D. (2006). eduGAIN: Federation Interoperation by Design. In T.-E. Research, & E. N. (TERENA) (Ed.), *TERENA Networking Conference 2006.* Catania.

Maali, F., & Erickson, J. (Eds.). (2014). Data Catalog Vocabulary (DCAT). *Data Catalog Vocabulary (DCAT)*. Retrieved June 10, 2018, from http://www.w3.org/TR/vocab-dcat/

Politze, M., Claus, F., Brenger, B., Yazdi, M. A., Heinrichs, B., & Schwarz, A. (2020). How to Manage IT Resources in Research Projects? Towards a Collaborative Scientific Integration Environment. In Y. Epelboin, M. Mennielli, A. Pacholak, P. Kähkipuro, G. Ferell, C. Diaz, . . . O. Tasala (Eds.), *European Journal of Higher Education IT 2020-1.* Paris, France. Retrieved from https://www.eunis.org/download/2020/EUNIS_2020_paper_39.pdf

Politze, M., Decker, B., & Eifert, T. (2017). pSTAIX – A Process-Aware Architecture to Support Research Processes. In M. Eibl, & M. Gaedke (Eds.), *INFORMATIK 2017: Digitale Kulturen* (pp. 1369–1380). Bonn, Germany: Köllen. doi:10.18420/in2017_137

Politze, M., Shakeel, Y., Hunke, S., Ost, P., Aversa, R., Heinrichs, B., & Lang, I. (2023, September). Long Term Interoperability of Distributed Research Data Infrastructures. *Proceedings of the Conference on Research Data Infrastructure, 1*. doi:10.52825/cordi.v1i.348

Speicher, S., Arwe, J., & Malhotra, A. (Eds.). (2015, February 26). Linked Data Platform 1.0. *Linked Data Platform 1.0*. Retrieved January 25, 2023, from https://www.w3.org/TR/ldp/

Weigel, T., Plale, B., Parsons, M., Zhou, G., Luo, Y., Schwardmann, U., . . . Kurakawa, K. (2018). RDA Recommendation on PID Kernel Information. *RDA Recommendation on PID Kernel Information*. (Research Data Alliance, Ed.) Research Data Alliance. doi:10.15497/RDA00031

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., . . . Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data, 3*. doi:10.1038/sdata.2016.18

7 Authors' biographies

Dr. Marius Politze is head of the department "Research Process and Data Management" at the IT Center of RWTH Aachen University. Before that he held various posts at the IT Center as software developer, software architect and as a teacher for scripting and programming languages. His research focuses on Semantic Web, Linked Data, and architectures for distributed and service-oriented systems in the area of research data management. *(CRediT: Project administration, Supervision, Conceptualization, Writing – original draft)*

Dr. Benedikt Heinrichs is a research associate and lead software developer at the IT Center of RWTH Aachen University since 2018. His research focuses on data provenance, metadata extraction, and similarity detection. He received his M.Sc. in Artificial Intelligence from Maastricht University in 2018. From 2013 until 2018, he worked at the IT Center as a software developer. *(CRediT: Conceptualization, Investigation, Software, Writing – review & editing)*

Sirieam Hunke is a research associate and software developer at the IT Center of RWTH Aachen University since 2022. She is part of the Task Area Materials Data Infrastructure within NFDI-MatWerk. Her Work focuses on federated and overarching architectures and implementation of the FAIR Digital Object concept. She received her M.Sc. in Energy Economy and Computer Science from FH Aachen in 2022. *(CRediT: Conceptualization, Investigation, Software, Writing – review & editing)*

Dr. Ilona Lang is the lead of the group "Data Management Platform Coscine" in the department "Research Process and Data Management" at the IT Center of RWTH Aachen University since 2022. From 2019 till 2021 she was Research Data Manager in different RDM projects at the KIM in at the University of Konstanz. Her work is focused on the ongoing strategic and user-orientated development of Coscine while coordinating the demands of different stakeholders. *(CRediT: Conceptualization, Investigation, Writing – review & editing)*

Dr. Thomas Eifert received his doctoral degree in solid state physics. Since 2013 he holds the role of the CTO at RWTH Aachen University's IT Center and is thus responsible for the technological strategy of the IT Center. His particular interests are the mutual dependencies of researchers' requirements and appropriate technical solutions, his teaching focuses on scalable IT. *(CRediT: Conceptualization, Writing – original draft, Funding acquisition, Project administration)*