The Triathlon Relay of Storage Space Provisioning

Ilona Lang [1*], Marcel Nellesen [1], Marius Politze [1]

[1] IT Center, RWTH Aachen Universiy;
[*] Corresponding author: lang@itc.rwth-aachen.de

The increasing demand for storage space in data-intensive sciences necessitates the establishment of cooperative information infrastructures among universities. This paper introduces the provisioning processes around DataStorage.nrw, a redundantly distributed cloud storage solution across four universities in North Rhine-Westphalia (NRW). Access to DataStorage.nrw is free for eligible researchers who meet specific criteria: they must be employed at a DH.NRW university, their storage needs must arise from a research project, and the data must adhere to FAIR principles. The allocation process resembles a triathlon relay, integrating various software disciplines to facilitate efficient resource provisioning. The workflow comprises three stages: Identity and Access Management (IAM) ensures secure user authentication; the RDM platform Coscine manages research metadata and project structures while streamlining compliance with FAIR principles; JARDS serves as the science-led review mechanism for storage applications, ensuring that resources align with funding objectives. By fostering cross-service interoperability and embracing agile processes, we aim to avoid isolated solutions and enhance collaboration among universities. Ultimately, our approach not only addresses immediate data management challenges but also provides a scalable framework that can serve as a blueprint for other universities seeking to establish similar cooperative infrastructures.

Keywords: Storage, IAM, Coscine, JARDS, FAIR Principles

Published in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger, and Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23918 (CC BY-SA 4.0).

1 Introduction

Data-intensive sciences require more and more storage space[1], which universities must provide. At the same time, the buffer calculations of individual universities lead to a sum of unused capacities in the many separate storage infrastructures. Part of the solution and national political demand is therefore to establish cooperative information infrastructures (Deutsche Forschungsgemeinschaft. Ausschuss für Wissenschaftliche Bibliotheken und Informationssysteme 2025) between universities. One of these information infrastructures is a redundantly distributed cloud storage for cooperative use – the DataStorage.nrw[2]. DataStorage.nrw comprises around 42 PB of usable storage space and is distributed across four locations at the Universities in Aachen, Cologne, Duisburg-Essen and Paderborn (see Figure 1). Storage space on the DataStorage is expected to be available from June 2025. The predecessor of DataStorage.nrw was the Research Data Storage (RDS), which was also redundantly distributed – but had a smaller storage capacity and was available to a smaller number of universities in NRW.

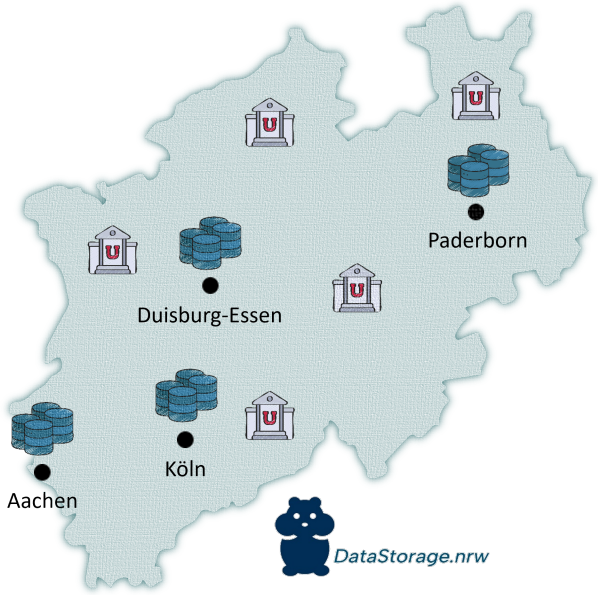

Figure 1: DataStorage.nrw – Map of the Locations.

Storage space on DataStorage.nrw is free of charge for researchers and a suitable storage location both for warm data during the active project phase and for cold data up to a maximum of 10 years after the end of the project. However, in addition to a cooperative use, the following criteria must be met: The researcher must be employed at a DH.NRW

1 https://www.kim.uni-konstanz.de/kim-news/news-detailseite/forschungsdateninventur-zeigt-grossen-speicherbedarf/; *Visited on March 31, 2025.*

2 https://www.land.nrw/pressemitteilung/riesen-wolke-macht-forschungsdaten-sicher; *Visited on March 31, 2025.*

university, their storage space requirement must have arisen within a research project, the storage be in compliance with the FAIR principles (Wilkinson et al. 2016). These three conditions must be checked as part of a scientific allocation procedure and require different tools to ensure that they are met. The review process is similar to a relay race in which different software disciplines come together to achieve the goal of allocating storage space – a triathlon relay of storage space provisioning. Table 1 gives an overview on all stages of the triathlon relay which will be described in the following chapters in more detail.

Table 1: Overview: Triathlon Relay of Storage Space Provisioning.

Stage	**Result**
1: Identity and Access Management	Cooperative use by all universities
2: Coscine – RDM Platform	Project management in a FAIR environment
3: JARDS – Software for Resource Allocation	Science-led review process
Finish line: DataStorage.nrw	Storage space allocation

2 First Stage: Identity and Access Management (IAM)

The first stage of the relay is the Identity and Access Management (IAM), it is important to know who the users are and what they can access. This is especially challenging with services that are available to a large variety of different institutions, and which supports cooperative use between universities. Additionally, an individual researcher might also change his university, go abroad or leave the university and go to the industry. In our platform we address this offering different login methods to the researchers. Firstly, they can use the DFN-AAI to login with their local identity provider, secondly social logins, such as ORCID are available for researchers that left the university, industry partners or international cooperation partners.

The Base4NFDI project IAM4NFDI[3] enables an easy connection of a service to one or more of the provided Community Authentication and Authorization Infrastructures (CAAI), e.g. the RegApp. Since these architectures are often complex, IAM4NFDI has regular calls for incubator projects. Within these technical questions and the connection of a service to the CAAIs are realized within a regular direct exchange between the service and the CAAI providers.

3 https://base4nfdi.de/projects/iam4nfdi; *Visited on March 31, 2025.*

One big benefit of the CAAIs is an easy connection of one service to many home organizations, without the need of a 1 to 1 communication between the different parties beforehand. The CAAI acts as an identity proxy in between the identity providers and the services. Therefore, identity provides only need to configure one identity proxy instead of multiple services and the services only need to configure one identity proxy instead of many identity providers. The utilized protocols SAML and OIDC are established and widely supported. This makes it also easier to transfer required information, which could include among others: Membership of a university, membership in a virtual organization, unified set of attributes and roles through IDM.nrw

This approach is that it is extremely scalable and represents the current state of art of identity and access management. As such it is also adopted by several federal and national projects such as the NFDI and NHR.

3 Second Stage: RDM Platform Coscine

The first relay handover is carried out from NFDI-Login to the RDM platform Coscine (re3data.org 2024). Coscine has been developed as open source software at RWTH Aachen University since 2018[4], has been in regular operation since 2023 and is used for the management of research (meta)data as well as for the allocation and provisioning of storage resources for research data. Coscine is developed according to the FAIR principles (Lang, Jansen, and Politze 2025) and implements interfaces for the so-called FAIR Digital Objects (Smedt, Koureas, and Wittenburg 2020). As part of Coscine.nrw, Coscine was established as a permanent service and thus made available to all DH.NRW universities. The relay transfer via the NFDI-Login provides Coscine with information on whether a person is an employee at a DH.NRW university. In addition, the login enables the registration of all universities within the DFN-AAI and, thanks to the ORCID option, also international researchers. Coscine itself has two main functions in the relay: To structure data storage in projects and ensure the implementation of a large part of the FAIR principles (Lang, Jansen, and Politze 2025).

Figure 2 shows the main workflow for researchers using Coscine. As described above, Coscine offers researchers a low-threshold login. After login, researchers create their projects using the web interface or via the Coscine API. Created projects and sub-projects are mapped in Coscine as a tree structure, enabling simple and hierarchical visualization. Project creators can invite all other project partners easily via e-mail and regulate their corresponding access via the project-related rights and roles concept. Coscine offers three different roles inside of projects: owners, members and guests. Based on the low-threshold login, external project partners can be invited to Coscine project, fostering collaborative co-operation across institutes and universities.

4 https://coscine.rwth-aachen.de/; *Visited on March 31, 2025.*

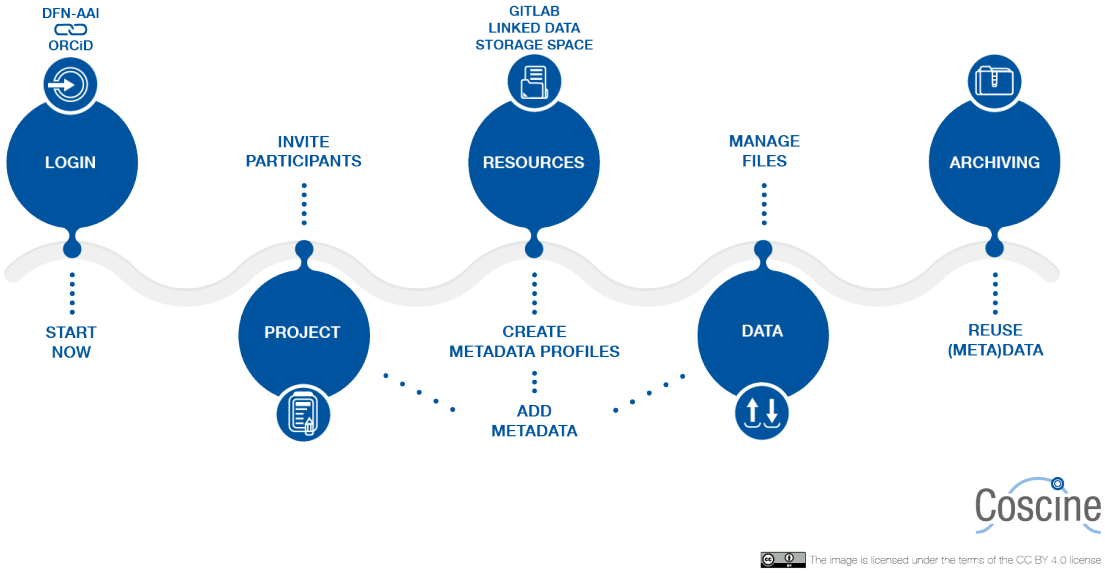

Figure 2: Coscine Workflow: Five Main Steps from Login to Archiving.

Research data is uploaded in Coscine at the resource level, where different resource types are available for each project depending on access rights and the selected login method of the respective project owner. For the DataStorage.nrw, the available storage space resource types are web resources (accessible only via the browser/API), S3 resources[5] (accessible via browser/API/S3) and WORM (write once, read many) resources (accessible via browser/API/S3). Apart from the DataStorage.nrw Coscine allows researchers to add other data sources of a project, for example via the GitLab resource type.

Coscine supports researchers by providing a rich metadata management environment that gives access to discipline specific metadata profiles. The metadata is automatically checked for completeness based on the refinements of each metadata profile. The metadata profiles are machine accessible based on their technical representation and validation via W3C standards RDF and SHACL as proposed by the AIMS Project (Grönewald et al. 2022). Moreover, Coscine implements the FAIR Digital Object Concept to represent stored data in concise digital entities (Politze et al. 2025; Heinrichs, Hunke, and Politze 2025). Coscine assigns for each project and resource (including data and metadata) a handle-based ePIC-PID (Kálmán, Kurzawe, and Schwardmann 2012). Individual files are described with metadata profiles, of which each RDF-triple includes PIDs leads to the described data by extending the handle URL. In summary, Coscine provides access to DataStorage.nrw while setting a project structure and enabling collaborative work in a FAIR environment.

As an intermediate stage, Coscine automatically allocates employees of DH.NRW universities 100 GB on DataStorage.nrw as web resource – a resource type with upload function

5 S3 (Simple Storage Service) is a simple object storage interface for data that supports the rapid retrieval of object data.

via the web interface and forced use of Coscine's metadata profiles. However, if researchers need more storage space and/or access via S3 resource, the relay needs to be passed on to JARDS[6], since Coscine can not guarantee the scientific character of a project and the linking to rich metadata in accordance with the FAIR principles (Lang, Jansen, and Politze 2025).

4 Third Stage: Science-led allocation procedure with JARDS

The third stage of the relay is the science-led review of the storage space applications. This stage is represented by the platform JARDS (Janetzko 2019). This abbreviation stands for Joint Application Review and Dispatch Service and was originally developed by the Jülich Super Computing Center for applications for computation time on the super computer. The RWTH Aachen University started an own instance in 2014 for its own high performance computer. Since then the platform was adapted by various universities and projects, e.g. the National High-performance Computing (NHR) and WestAI.

In 2022 this platform was adopted to support applications for storage space in Coscine. Multiple application types are supported based on the resource types that is requested. While less configurable DataStorage.nrw resources such as the *web* resource only require a minimal application more specialized resources such as *S3* and *Write Once Read Many (WORM)*, require a detailed description of the metadata management, the stored data and the data flows.

Figure 3 shows the main workflow from the DataStorage.nrw application to review using JARDS. After the researcher has submitted the application for storage capacity by filling out all required fields regarding the project and data handling, the review process is initiated. As a first step, the proposal is formally evaluated: This means that it is checked first, if the applicant is eligible to request a storage capacity and second, whether the answers are complete and contain all needed information. This step is conducted by members of the universities' RDM team, and the formal evaluation typically takes between one or two days.

Once the evaluation is done, in the next stage the technical and scientific review is performed (Figure 3). Within the technical review, staff of the local RDM team will review the application for technical feasibility with special focus on the proposed data and metadata management (see *Criteria for Technical Evaluation* below). In case of problems or questions, the principal investigator (PI) and/or the person of contact (PC) of the project are contacted to provide the missing information or to adjust the plan to ensure good research data management practices. This is roughly equivalent to a data management

6 https://jards.coscine.de/; *Visited on March 31, 2025.*

plan review. Usually, this step takes about one week. After a successful first technical review by the home institution, a second technical review is carried out by an independent university from DH.NRW, to ensure a fair review process. This step takes about another week.

Based on the amount of storage space that is requested, the review process can be extended with a third step: the scientific review. When researchers require more than 125 TB, a single-blind review of the project application by up to three independent domain scientists from German universities or other research facilities is performed. These domain scientists can suggest adjustments to both the envisioned process and the requested storage space. Because of these external dependencies, this process takes between four and six weeks for DataStorage.nrw applications for *web* resources and up to three months for applications for *S3* and *WORM* resources.

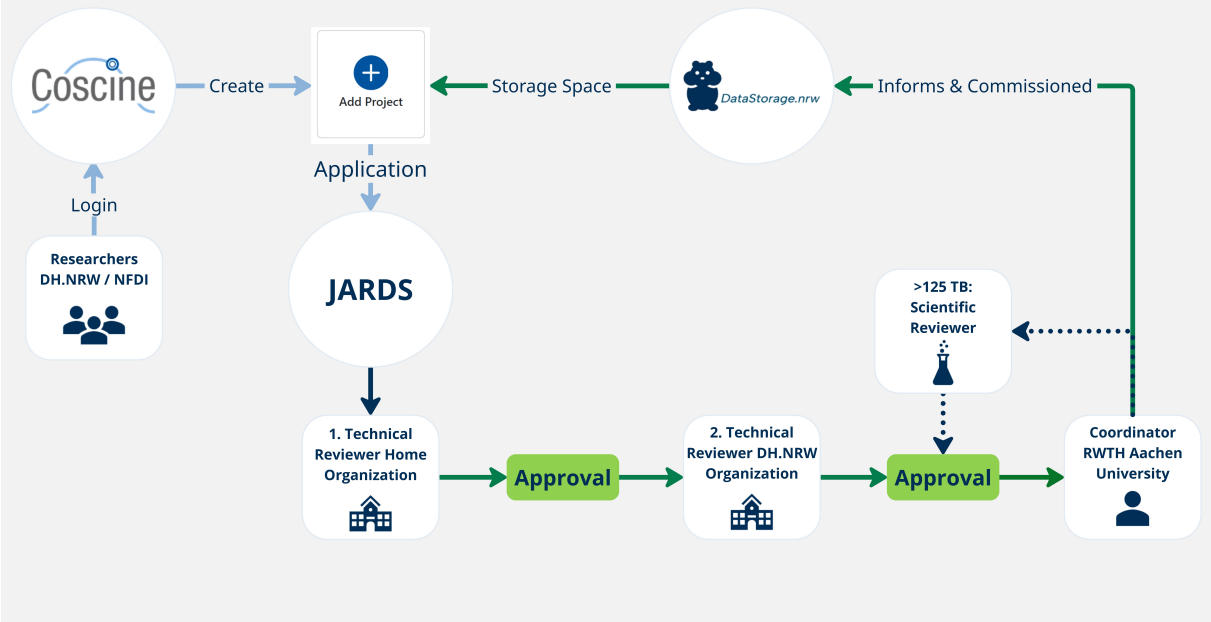

Figure 3: JARDS Workflow: The Main Steps Application, Review Process and Storage Allocation.

The technical evaluation of storage applications has the main objective to sensitize researchers for the FAIR principles and research data management in general. There are three main criteria that are part of the evaluation:

- **Project Context:** Is the application based on a research project?
- **Amount of Requested Storage:** Is the requested storage space suitable for the described project?
- **Metadata:** Is the data annotated with detailed metadata? Is metadata added in a consistent and reliable way? Is the structure and the findability of the metadata ensured?

The first question is aimed at ensuring the project structure. For example, it is not permitted to store photos from the last company trip on DataStorage.nrw (background: no research project and no research data). Likewise, it is not permitted to store research data without a specific project reference, as otherwise no data life cycle can be ensured (for example, there is no clear project end date). The second criterion is to ensure that the scope of the requested storage roughly matches the described project – a 125 TB application from a project with pure literature research, for example, would be questioned. The third criterion is aimed at the area of FAIR principles, which cannot be automatically ensured by Coscine if using S3 resources. This concerns the researchers' handling of metadata. In most cases, this is also the question area that generates the greatest consulting effort. The reviewers often provide advice on subject-specific metadata profiles – combined with the recommendation to use the metadata management tools provided by Coscine.

5 Finish Line and Outlook

After successfully completing all three relay stages, the researchers reach the finish line and receive their storage space on DataStorage.nrw. They can now manage this storage space independently in their Coscine projects.

With the new storage provision process presented here, we are moving from unstructured data volumes to clearly defined research project units. This transition and the focus on the FAIR principles is necessary to enable a regulated data life cycle and at the same time provide a large number of different universities with storage space. The first stage of IAM is crucial for collaboration – if not every potential project partner can log in, just a new isolated solution for data storage is created. The merging of different accounts via different login options has so far generated a lot of effort in the development of Coscine – thanks to the help of IAM4NFDI, we could remedy this situation.

The second stage in Coscine is crucial for the clearly defined research project units in a FAIR environment. Our aim is always to make as little work as possible for the researchers by automatically fulfilling most of the requirements of the FAIR principles through Coscine.

The third stage by JARDS covers a key feature of the DataStorage.nrw operating concept which is the demand for a science-led management. JARDS helps to ensure that DataStorage.nrw resources are used in line with the funding objectives and offers different features for a collaborative review process. Here, RDM personnel from NRW has to take on an essential role as "Gate Keepers" for storage space. A critical aspect of this gate keeping is the joint revision of technical evaluation criteria for storage applications to ensure that our review processes remain relevant and effective. Along with an update to the content, we are planning closer technical integration between Coscine and JARDS (e.g. automated reviewer and storage assignment). This should increase the technical

efficiency and sustainability of the processes. Another challenge is, that we only perform an initial curation during deposition, with no regular review of compliance with the criteria. In the future, JARDS could also play an essential role in a regular curation via its functions for projects.

Overall, we try to avoid isolated solutions. It is crucial that we consider cross-service interoperability in our processes. This approach provides us with greater flexibility concerning the underlying hardware while ensuring seamless integration across RDM services. The developments of the open source platforms Coscine and JARDS relay heavily on agile processes that facilitate regular adjustments based on RDM personal and researcher feedback and their changing needs. This requires significant training and communication efforts within the service management and development teams. One part of our solution is to focus on scalability and transparency across all participating processes in the 'Triathlon Relay of Storage Space Provisioning' to ensure smooth operation and adaptation over time.

Taken together, we aim for the regular further development of all components by engaging all parties involved. Our strategy involves linking services from NFDI, HPC, and storage infrastructure into a cohesive process. The advantages include scalability, cost reduction, and increased acceptance among researchers. However, we must also be mindful of potential disadvantages such as dependence on multiple services and the associated synchronization efforts. Another pressing issue of DataStorage.nrw is balancing storage duration with the lifespan of hardware. It is essential to recognize that cooperative infrastructures do not automatically equate to sustainability; ongoing financing is crucial for maintaining hardware. In this respect, we remain dependent on a binding political commitment to ensure the storage of data for at least ten years.

In conclusion, by embracing these challenges and fostering collaboration among all stakeholders, we are not only enhancing our data storage solutions but also paving the way for a sustainable and efficient research environment that aligns with the evolving needs of academia. Our commitment to the FAIR principles and cross-service interoperability help researchers to focus on their projects while benefiting from a reliable and well-structured data management system. The processes described here can serve as a template for the implementation of other cooperative information infrastructures between universities.

Acknowledgements

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). 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). The IAM4NFDI project is partially funded by DFG as part of NFDI (grant number 521466146). The work

on Coscine was partially supported with resources granted by DH.NRW for the project Coscine.nrw by Ministerium für Kultur und Wissenschaft des Landes Nordrhein Westfalen, by NFDI4Ing, funded by Deutsche Forschungsgemeinschaft (DFG) under project number 442146713 and NFDI-MatWerk, funded by Deutsche Forschungsgemeinschaft (DFG) under project number 460247524.

Authorship Contributions

Please specify the contributions of each author to the manuscript.

- Ilona Lang: Conceptualization, Methodology, Writing
- Marcel Nellesen: Conceptualization, Methodology, Writing
- Marius Politze: Conceptualization

Conflict of Interest

The authors declare no conflict of interest.

Bibliography

Deutsche Forschungsgemeinschaft. Ausschuss für Wissenschaftliche Bibliotheken und Informationssysteme. 2025. *Digitale Forschungspraxis und kooperative Informationsinfrastrukturen. Ein Diskussionspapier der Deutschen Forschungsgemeinschaft zu Förderung und Finanzierung wissenschaftlicher Informationsinfrastrukturen.* Zenodo. https://doi.org/10.5281/zenodo.14621979.

Grönewald, Matthias, Patrick Mund, Matthias Bodenbrenner, Marc Fuhrmans, Benedikt Heinrichs, Matthias S. Müller, Peter F. Pelz, et al. 2022. "Mit AIMS zu einem Metadatenmanagement 4.0: FAIRe Forschungsdaten benötigen interoperable Metadaten". In *E-Science-Tage 2021: Share Your Research Data: Share Your Research Data,* edited by Vincent Heuveline and Nina Bisheh, 91–104. Heidelberg: heiBOOKS. ISBN: 978-3-948083-54-0. https://doi.org/10.11588/heibooks.979.c13721.

Heinrichs, Benedikt, Sirieam Hunke, and Marius Politze. 2025. "Combining FAIR Digital Object Implementation Concepts in a Real-Life Application". *Open Conference Proceedings* 5. https://doi.org/10.52825/ocp.v5i.1173.

Janetzko, Florian. 2019. “JARDS Ein Softwarewerkzeug zur Handhabung von Ressourcenvergabeprozessen”. In *ZKI-AK Supercomputing Herbsttagung,* Berlin, Germany. https://juser.fz-juelich.de/record/868324.

Kálmán, Tibor, Daniel Kurzawe, and Ulrich Schwardmann. 2012. “European Persistent Identifier Consortium – PIDs für die Wissenschaft”. In *Langzeitarchivierung von Forschungsdaten: Standards und disziplinspezifische Lösungen,* edited by Reinhard Altenhöner and Claudia Oellers, pages 151–164. Berlin, Germany: Scivero Verl. ISBN: 978-3-944417-00-4.

Lang, Ilona, Katja Jansen, and Marius Politze. 2025. *Implementing the FAIR Principles with Coscine.* Zenodo. https://doi.org/10.5281/zenodo.14886335.

Politze, Marius, Benedikt Heinrichs, Sirieam Hunke, Ilona Lang, and Thomas Eifert. 2025. “FAIR Digital Objects: FAIRtilizer for the Digital Harvest”. In *EPiC Series in Computing,* 105:284–274. EasyChair. https://doi.org/10.29007/hfzk.

re3data.org. 2024. *Coscine.* re3data.org – Registry of Research Data Repositories. https://doi.org/10.17616/R31NJNJZ.

Smedt, Koenraad de, Dimitris Koureas, and Peter Wittenburg. 2020. “FAIR Digital Objects for Science: From Data Pieces to Actionable Knowledge Units”. PII: publications8020021, *Publications* 8 (2): 21. https://doi.org/10.3390/publications8020021.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersbergand, Gabrielle Appleton, Myles Axtonand, Arie Baakand, Niklas Blombergand, et al. 2016. “The FAIR Guiding Principles for scientific data management and stewardship”. *Scientific data* 3 (1): 1–9. https://doi.org/10.1038/sdata.2016.18.