Carrots and Sticks: Motivating with Storage for Good RDM – Science Led Allocation of Research Data Storage Resources within an Integrated RDM System

Ilona Lang, Marcel Nellesen, Lukas C. Bossert, Marius Politze

IT Center, RWTH Aachen University

Storage space is valuable and there are many researchers who need to store their research data (also demanded by the Good Scientific Practice (GSP); Deutsche Forschungsgemeinschaft e.V. 2019). Most existing storage distribution systems are ad-hoc, require (internal) transfer of funds, or do not scale on institutional or even national level. Most importantly, the value for the scientific community often remains unaddressed. Starting with our existing data management platform Coscine we adapted the Joint Application Review and Dispatch Service (JARDS; Janetzko 2019), a tool already utilized within many computing centers within Germany to handle applications for computing time. Hence, our aim is to unify applications for scientific IT resources and lighten the process of formalities management.

1 Introduction

The structured organization of research data is eminent to research projects. And since metadata are more and more required to fulfill the requirements of e.g., FAIR principles (Wilkinson et al. 2016) and/or GSP, researchers are confronted not only with the task to find a suitable storage system for their data along with the metadata, but also they need to find a system with enough storage capacity for ongoing and finalized projects. At the RWTH Aachen University, we support researchers with the research data management platform Coscine (Politze et al. 2020). Coscine enables researchers to store their data along with all needed and demanded customized metadata. It further provides sufficient storage capacity in a secured and through Coscine easily accessible and manageable way. To achieve this, Coscine combines (decentral) data storage systems with a metadata management (Schmitz and Politze 2018; Politze et al. 2020). Technically it leverages persistent identifier (PID; Kálmán, Kurzawe, and Schwardmann 2012; Krämer, Politze, and Schmitz 2016) and linked data technologies on multiple levels: projects, storage

Publiziert in: Vincent Heuveline, Nina Bisheh und Philipp Kling (Hg.): E-Science-Tage 2023. Empower Your Research – Preserve Your Data. Heidelberg: heiBOOKS, 2023. DOI: https://doi.org/10.11588/heibooks.1288.c18071

resources and individual files and applies the FAIR Digital Object (FDO; Smedt, Koureas, and Wittenburg 2020) concept and Data Catalog Vocabulary (DCAT; Maali and Erickson 2014, cf. Figure 1). One of the core storage systems behind Coscine is Research Data Storage (RDS; Eifert, Claus, and Lopez 2018), a geo-redundant object storage system that is provided by a consortium of universities for all researchers within the federal state of North Rhine-Westphalia and their collaboration partners within the National Research Data Infrastructure (NFDI).

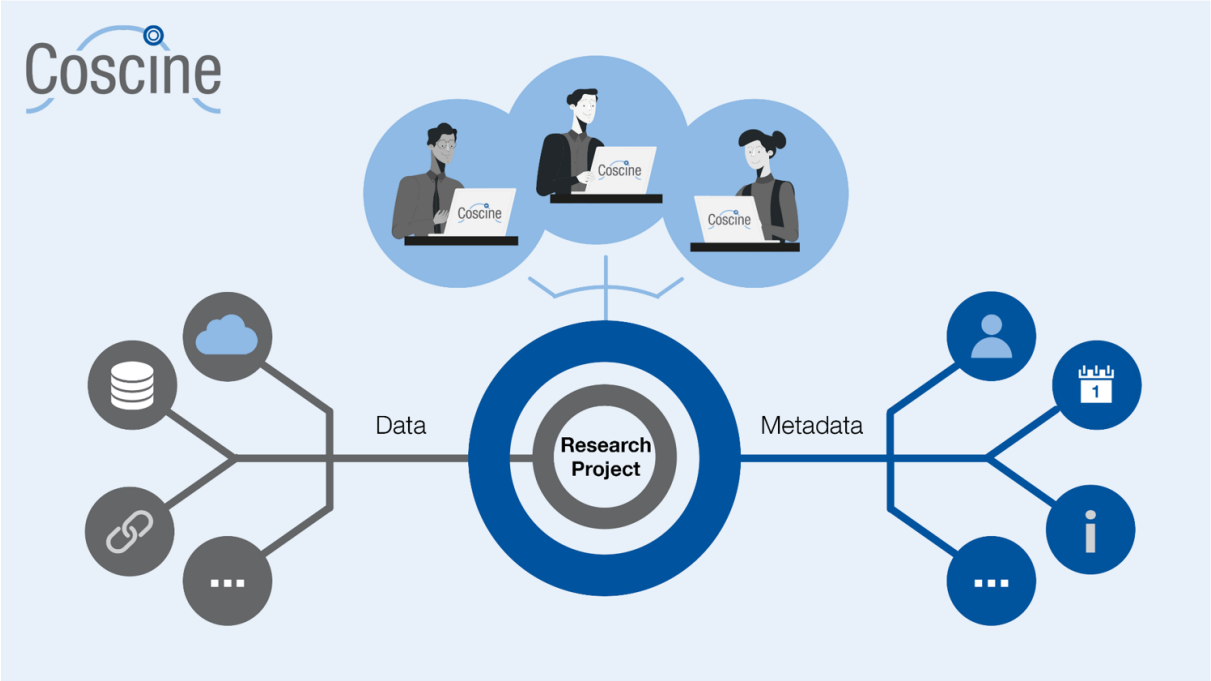

Figure 1: Data and metadata in Coscine.

Depending on the project and the data, the required amount of storage capacity varies strongly. Furthermore, the knowledge on research data management is also individual, and therefore research data is handled, organized, and annotated differently. To support researchers in their needs and at the same time ensure that the data and the corresponding metadata is managed correctly, the research data management (RDM) team not only consults and offers workshops, and when it comes to the request for storage capacity we created a digital process that guides researchers through the steps of describing their research project and how they intend to manage the corresponding research data. Hence, the carrots and sticks metaphor: rewarding good RDM practices with access to data storage systems. The required application process is conducted by the use of JARDS. JARDS is already used in the context of applying for computer time at various high performance computing (HPC) systems in Germany.

2 Workflow

As a prerequisite, researchers need to create a project in Coscine. At this point they already have to provide certain meta information about the project (description, time frame, collaborating people and institutions etc.). Having collected the meta information for a research project grants a limited default storage quota that can be used directly. When it comes to extend this default storage capacity for a project, researchers can follow a science led application process close to the peer review of a research contribution. Researchers will find a documentation on the pages of Coscine how to proceed[1].

In the following sections, we will describe the workflow of the required steps for the application and review process. After describing the preparation and submission steps, we will explain the formal evaluation and as well technical as scientific review of the application. Further on, we talk about the resource allocation and monitoring steps and how the storage capacity is included in the reporting.

2.1 Project preparation and proposal submission

In Coscine researchers can create various forms of resources in which the data is stored. The resource types not only differentiate in how data is mainly uploaded and annotated by metadata (RDS-Web: via a web interface or a custom Application Programming Interface (API) that enforces metadata quality, RDS-S3: via the widely used S3 protocol) but also on the persistent integrity of once uploaded data (RDS-WORM: write-once-read-many-storage that does not allow changes once a file is stored).

In JARDS the different resource types are represented since they require different information from the researcher. For resource types that ensure correct handling of metadata and other good RDM practices the form is simpler (especially in the case of RDS-Web), the more specific the requirements of the researchers are the more information they must provide. The most information currently is required for the resource type RDS-WORM, since incorrect use of the resource will block valuable storage space for 10 or more years.

After researchers have identified a resource type that matches their requirements based on the flow diagram that is shown in Figure 2, the application process is initiated. Through the system, researchers can file an arbitrary amount of storage applications for one or multiple projects within Coscine, as they assume being appropriate for their scientific workflow. JARDS offers an overview of the current status of these applications (cf. Figure 3). For getting some context about the project and contact information, a very first question demands the project title, description and PI or PC (cf. Figure 4).

The application workflow ensures that on the one hand the application provides scientific value and on the other hand that there is at least a basic data management plan (DMP). As such, important questions are what kind of data is created or processed within this project. Are there any special requirements for the data that need to be considered, e.g.,

1 `https://docs.coscine.de/en/projects/storage/`; *Last accessed on May 15th, 2023.*

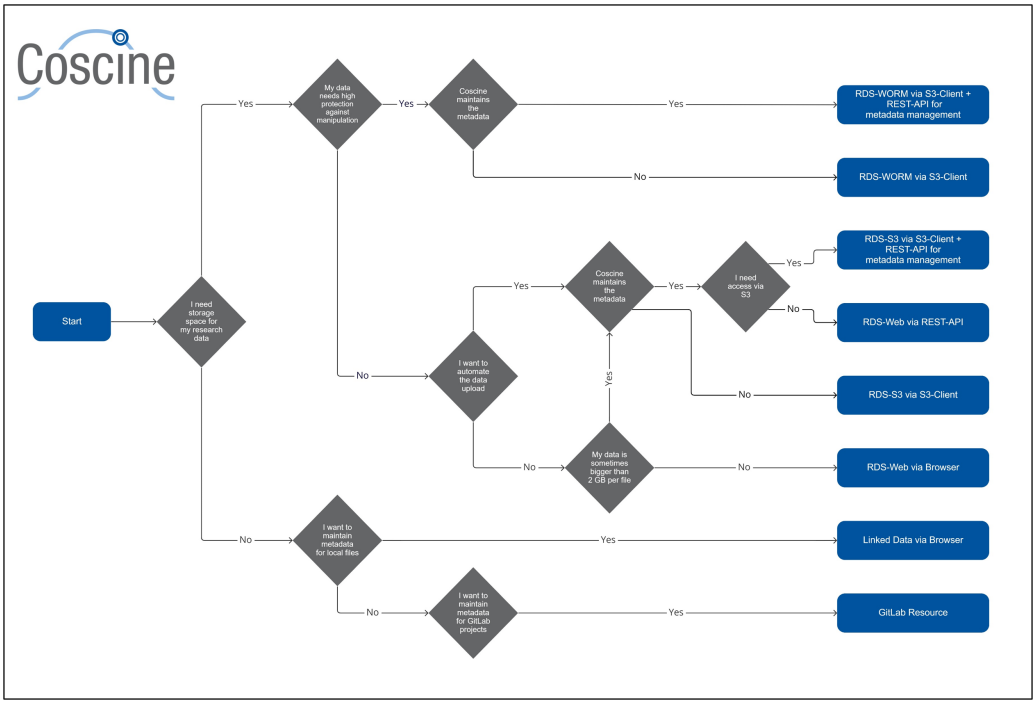

Figure 2: Application selection.

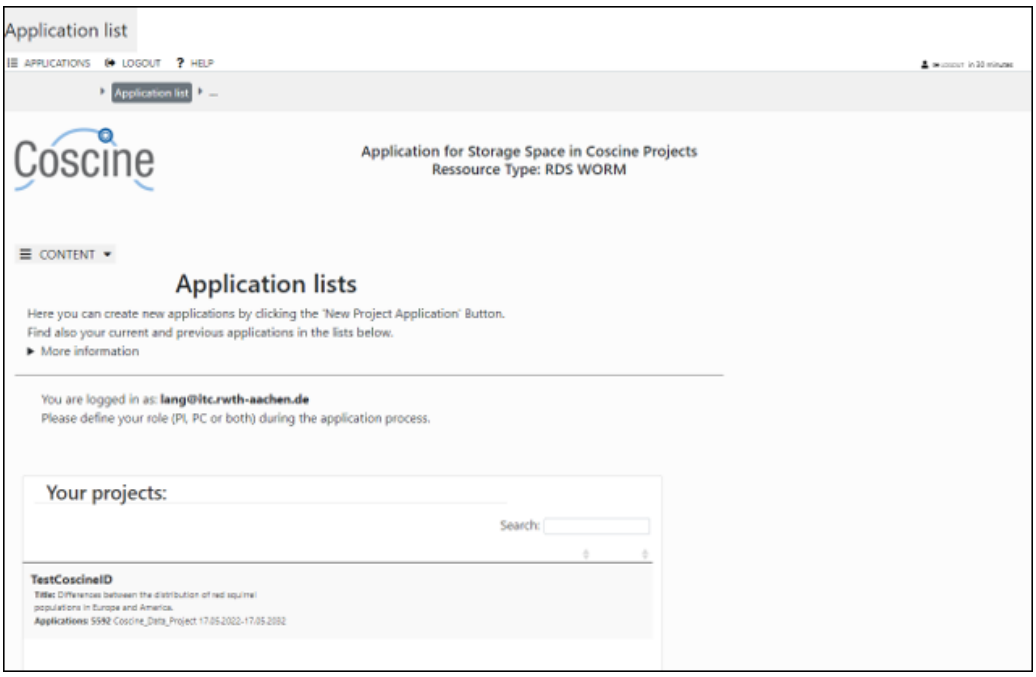

Figure 3: Applying for storage space in JARDS: Application Creation.

data privacy concerns, ensuring the raw data cannot be altered or the usage of distributed data sources. A core part is about the internal structure of the research data and how researchers intend to handle or organize it.

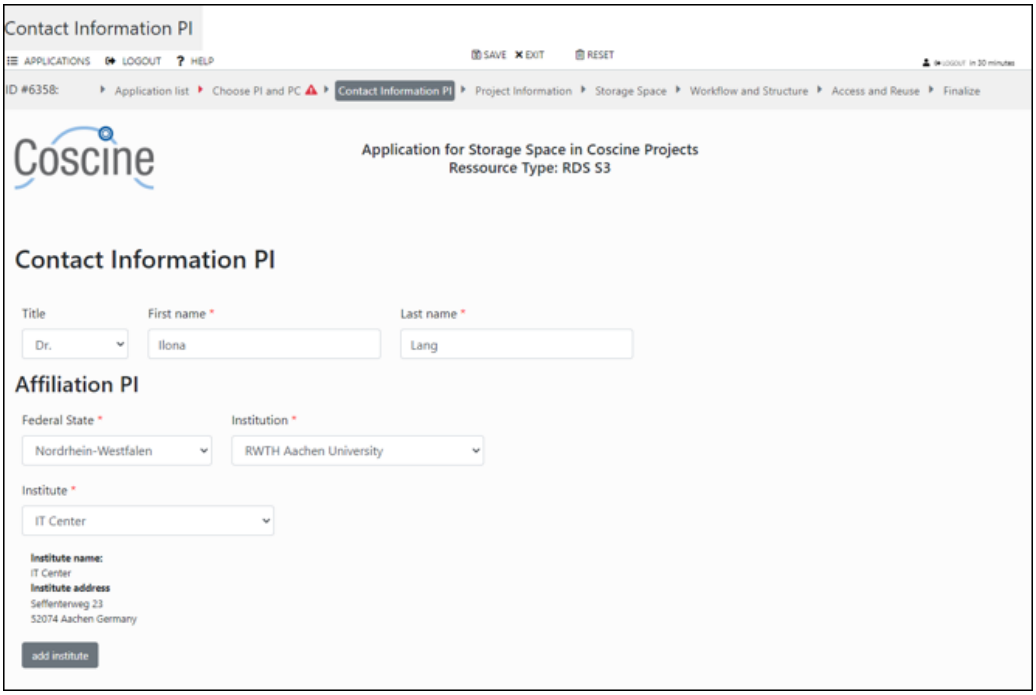

Figure 4: Applying for storage space in JARDS: PI Information.

The crucial question is about the amount of storage capacity. The default quota for an RDS-Web resource is 100 GB. For the resource type RDS-S3 and RDS-WORM, there is no default quota. Researchers can name any figure, but it needs to be plausible in respect of the described project and handling of data.

All the given information are part of the technical, and in case of the request of more than 125 TB storage capacity, also scientific review.

2.2 Formal evaluation, technical and scientific review

After the researcher has submitted the formal request 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 (cf. Figure 5). 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. 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 plan review. Usually, this step takes about one week.

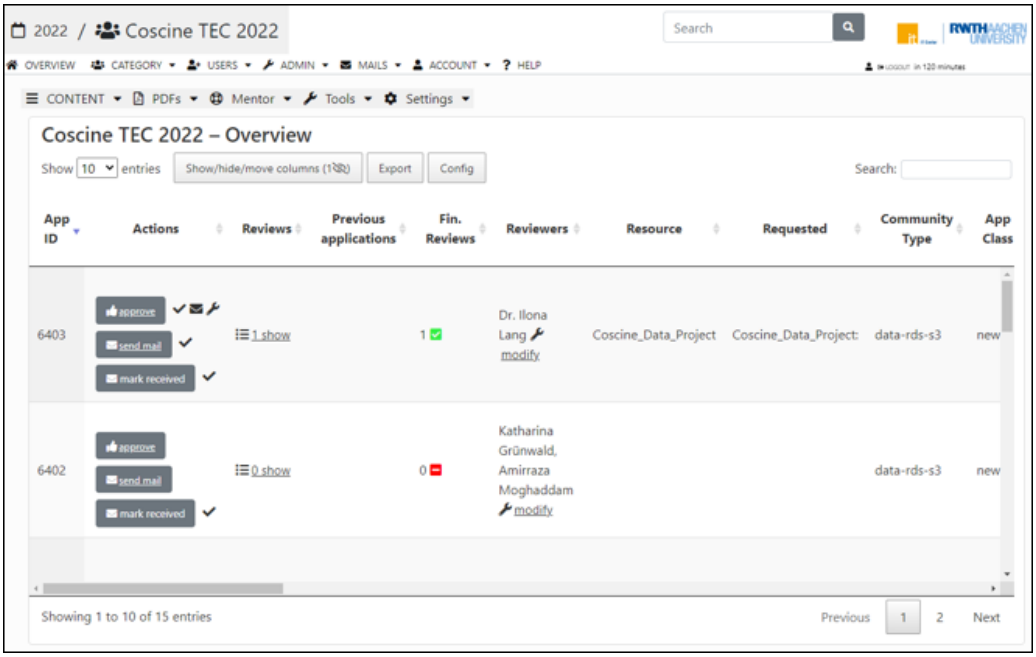

Figure 5: Review component.

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 applications for RDS-Web and up to three months for applications for RDS-S3 and RDS-WORM.

2.3 Resource allocation and monitoring

After the review process is completed, the requests will be either rejected or approved. In the latter case, quota will be granted. In case special configurations were requested, a training or counselling is offered to the applicant to ensure correct usage of the system. This approach offers a unique possibility for the universities' RDM team to get into contact with the heavy users of data storage infrastructures and to increase digital literacy and competences in a targeted manner.

When the review process is finalized, the application is approved, the requested resources are assigned within Coscine to the project of the applicant. Since the review process can take longer for larger applications, a preliminary initial quota can be provided for certain categories. This enables the researches to set up their workflows with the storage systems while waiting for the final review of their application. After the application was approved, the quota will be extended and the size of the initially created resources can be easily adjusted within Coscine (cf. Figure 6). The PI/PC can add further users to their projects

within Coscine at any time and can also monitor their available and utilized quota at any time.

In case the originally requested resources are not sufficient, an extension can be requested. The original application can be used as a base for the application for an extension, which will then be reviewed as described above. JARDS also provides an additional option: if small amounts of additional quota are required, a project can be extended once to grant an additional 25 % of the original quota. This small extension does not require a complete review process.

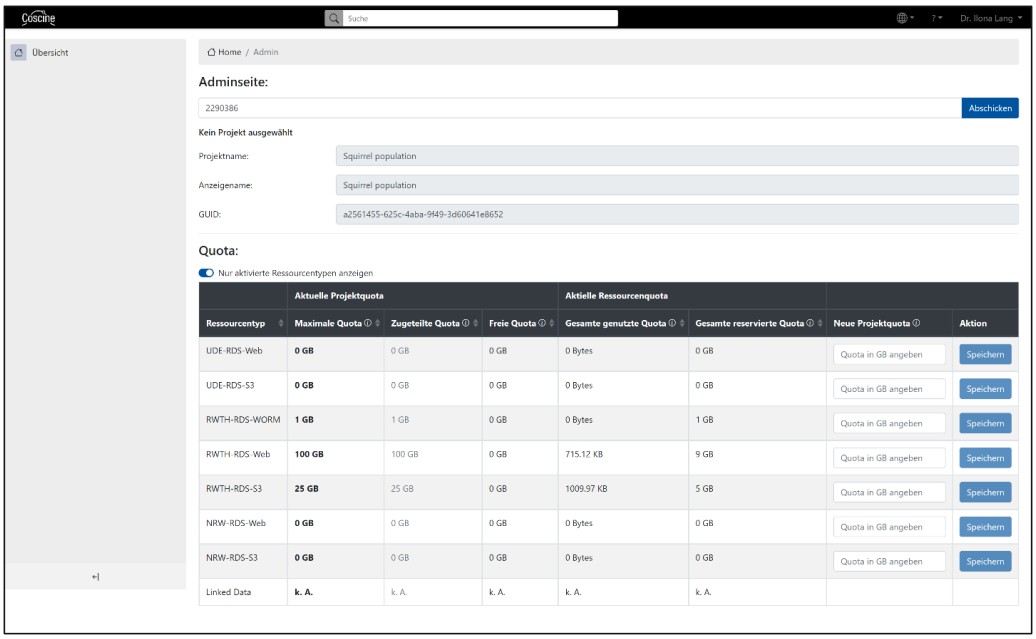

Figure 6: Quota management.

2.4 Reporting

JARDS also provides the option for the RDM staff to manage existing projects within the project component. Within this component the users can see all their approved projects, and the granted resources. There is an option for system operators to automatically report the amount of utilized resources, so the PI can monitor the still available resources and request more storage space if required. The component also offers different options for operators and managers of the storage system, e.g. there are regular status reports and a final report can be requested from the researchers. The researchers are contacted through mail and can upload these reports within JARDS. In addition, publications that were created as parts of the research project can be entered within the component as well.

3 Conclusion and outlook

The workflow presented in the previous section can easily be extended to include different resource types. This can be other storage systems, computation time on a high-performance computing system, or any other IT resource. Through the science led review process, all applicants are treated equally throughout the entire process. This not only ensures a quality standard but could also enable comparability between different applications, in case a strongly limited resource is managed with the system. The system is scalable and the number of operators and reviewers for each resource can be adjusted according to the requests. Additionally, this allows the allocation and provision of statewide available storage resources, such as RDS, according to uniform criteria by the science led management concept within national service offerings like Coscine.nrw[2]. In addition to management, this supports the storage of research data according to the FAIR principles. This improves participation opportunities of smaller universities in these scientific (storage) infrastructures and thus increases the economic efficiency of the invested resources in the long term.

The presented approach forces researchers to think about their data and the corresponding metadata from the start of the project. It also provides a unique opportunity for the universities' RDM team to reach out to heavy data users and supply them with targeted information about the systems used, or to build tailored offers to enhance digital literacy. The process has several similarities to the submission and review of scientific papers, and therefore is familiar to the researchers. Another advantage is that many researchers are already familiar with the utilized software and its functions, since they use the same software to apply for computing time projects on many HPC clusters in Germany. This allows an easier adaptation of the software for the researchers and can give HPC centers the possibility to combine applications for computing time projects and data projects.

Acknowledgements

The work was partially 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.

References

Deutsche Forschungsgemeinschaft e.V. 2019. *Guidelines for Safeguarding Good Research Practice: Code of Conduct.* Bonn, Germany. https://www.dfg.de/download/pdf/foerderung/rechtliche_rahmenbedingungen/gute_wissenschaftliche_praxis/kodex_gwp_en.pdf.

2 https://www.dh.nrw/kooperationen/Coscine.nrw-100; *Last accessed on May 15th, 2023.*

Eifert, Thomas, Florian Claus, and Ania Lopez. 2018. *Research Data Storage (RDS): Verteilte Speicherinfrastruktur für Forschungsdatenmanagement: Gemeinsamer Antrag (öffentliche Fassung) im DFG-Programm "Großgeräte der Länder": RWTH Aachen University (Konsortialführer), Fachhochschule Aachen, Ruhr-Universität Bochum, Technische Universität Dortmund, Universität Duisburg-Essen, Universität zu Köln.* Technical report. DOI: https://doi.org/10.18154/RWTH-2021-04541.

Janetzko, Florian. 2019. "JARDS Ein Softwarewerkzeug zur Handhabung von Ressourcenvergabeprozessen". In *ZKI-AK Supercomputing Herbsttagung.* 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.

Krämer, Florian, Marius Politze, and Dominik Schmitz. 2016. *Empowering the Usage of Persistent Identifiers (PID) in Local Research Processes by Providing a Service and Integration Infrastructure.* In collaboration with RD Alliance. Garching, Germany.

Maali, Fadi, and John Erickson, editors. 2014. *Data Catalog Vocabulary (DCAT).* W3C. Visited on June 10, 2018. http://www.w3.org/TR/vocab-dcat/.

Politze, Marius, Florian Claus, Bela Darius Brenger, Mohammad Amin Yazdi, Benedikt Paul Anton Heinrichs, and Annett Schwarz. 2020. "How to Manage IT Resources in Research Projects? Towards a Collaborative Scientific Integration Environment". *European journal of higher education IT* 1 (2020/1): 5. ISSN: 2519-1764. DOI: https://doi.org/10.18154/RWTH-2020-11948. https://publications.rwth-aachen.de/record/808269.

Schmitz, Dominik, and Marius Politze. 2018. "Forschungsdaten managen – Bausteine für eine dezentrale, forschungsnahe Unterstützung". *o-bib. Das offene Bibliotheksjournal* 5 (3): 76–91. DOI: https://doi.org/10.5282/o-bib/2018H3S76-91.

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. DOI: 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. DOI: https://doi.org/10.1038/sdata.2016.18.