Scale=MatchLowercase []Ligatures=TeX,Scale=1 LiberationSerif[Path = ./fonts/liberation/, Extension = .ttf, UprightFont = *-Regular, BoldFont = *-Bold, ItalicFont = *-Italic, BoldItalicFont = *-BoldItalic] LiberationSans[Path = ./fonts/liberation/, Extension = .ttf, UprightFont = *-Regular, BoldFont = *-Bold, ItalicFont = *-Italic, BoldItalicFont = *-BoldItalic] LiberationMono[Path = ./fonts/liberation/, Extension = .ttf, UprightFont = *-Regular, BoldFont = *-Bold, ItalicFont = *-Italic, BoldItalicFont = *-BoldItalic]

RESEARCH ARTICLE

Coscine – FAIR play integrated right from the start

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

1. IT Center, RWTH Aachen University, Aachen.

Abstract. Max. 150 words. The background of the research topic : Lorem ipsum dolor sit amet, consectetur adipiscing elite.

Keywords:
Inggrid, Data

Data availability:

Software availability:
Software can be found here:
coscine.rwth-aachen.de/

1 Introduction

For many researchers, whether from engineering sciences or other fields, an involvement with the FAIR Guiding Principles [**Wilkinson.2016**] does not begin until the publication of an article and the sometimes-obligatory transfer of the research data to a repository. At this point, a significant amount of valuable information about the research project is often already lost. Therefore, only a fraction of the data (and metadata) collected during a research project is ever published.

But even if researchers try to follow the FAIR principles during their whole data life cycle, it is a big challenge to find a service that offers solutions for all project-related data types (e.g., managing code, collaborative work, multiple large files). Therefore, researchers typically employ a broad spectrum of IT service infrastructures for their projects that range from local to centralized, federated and external IT service providers. Central applications like Radar [**Kraft.2016**] or MASi [**Grunzke.2019**] are less specific and address a wider community with more generic **RDM!** (**RDM!**) workflows. External "clouds" like Zenodo, Figshare or **OSF!** (**OSF!**) support basic **RDM!** workflows like citation or persistent identification. By far most prominent are generic "clouds" like the Owncloud-based Sciebo [**Vogl.2015**], Dropbox, Google Drive or GitLab to store and manage data, however, these options usually lack in support of **RDM!** workflows or policies.

Taken together, the situation nowadays often leads to a fragmentation of research data among a multitude of service providers with varying (if any) levels of maturity with respect to FAIR **RDM!**. Moreover, the amount of service providers makes it hard for researchers to keep an overview over the entirety of data related to a research project.

Thus, a software solution is needed to get all research data under one roof while supporting the FAIR principles. Based on the focus on engineering at RWTH Aachen University and the associated high volume of research data, initial analyses and developments towards such a software solution were started at the **RDM!** team of the IT Center in

2018 . Two options were analyzed: 1. develop a data management system that replaces all existing services or 2. develop a data management system that adds a "FAIR" layer to already established services. The first option would require an enormous amount of human resources and the willingness of researchers to give up all previously used services for a new system. Research in the field of **RDM!** shows, however, that software development (especially in the public sector) is confronted with low human resources (proof XYZ) and the willingness to change established software among researchers is low (proof XYZ). Both challenges make the development of a data management system that replaces all existing services an unattainable goal in the near future. The second option thus has two direct advantages: 1. the data management system does not have to cover all the functions of already established services but can focus entirely on adding features for compliance with the FAIR principles and 2. researchers can use all their established services and still get access from one platform.

To create such a data management system, the research data management platform Coscine was developed at the IT Center of the RWTH Aachen University. Since 2020, the development is further supported by two consortia of the **NFDI!** (**NFDI!**): NFDI4Ing [**Schmitt.2020**] and NFDI-MatWerk [**Eberl.2021**]. These consortia aim to develop **RDM!** solutions that, at best, can be applied to other disciplines as well. For the engineering sciences, the NFDI4Ing was founded to develop, disseminate, standardize and provide methods and services to make engineering research data FAIR (https://nfdi4ing.de/about-us/).

In this paper, we show which features Coscine provides for researchers and how they support the FAIR principles - from the initial collection of data to its subsequent reuse.

2 Core Features of Coscine

Coscine is a platform for the management, storage and archiving of research data and metadata generated in the context of research projects. The service is designed to support researchers in **RDM!** and the preservation of **GSP!** (**GSP!**). Specifically, Coscine offers researchers the following core features:

Integration By integrating various already established services, so-called resources, researchers can see and manage all project data in one place via the Coscine web interface or the Coscine API. Currently, the resources of the **RDS!** (**RDS!**) and Linked Data are integrated. Planned for early 2023 is the integration of GitLab. For the end of 2023 cloud applications such as Sciebo and Nextcloud shall be integrated. Based on customer requests or market changes, additional resources can be continuously added or others replaced.

Storage Space Coscine provides access to storage space on the **RDS!**. By default, employees of participating universities receive 100 GB of storage space per project for their research data, which they can distribute among several **RDS!**-Web resources. For large amounts of data, more storage space can be requested. It is also possible to request **RDS-S3!** (**RDS-S3!**) resources to interact directly with the **S3!** (**S3!**) buckets.

Collaboration Coscine allows access for all internal and external members of a research project. Users can log in as a member of a participating organization via Shibboleth or as an external person via their **ORCID!** (**ORCID!**). Project members can be added to projects in a low-threshold way via their email, enabling easy collaborations.

Metadata The use of Coscine involves three levels of metadata: at the project, resource, and data level. Adding metadata at the project and resources level is mandatory and the necessary fields are standardized for all users and disciplines. At the data level users can choose between different application profiles to optimally describe their research data. Individual application profiles can be created using the integrated AIMS application profile generator. All metadata are captured according to flexibly definable schemas that follow **RDF!** (**RDF!**), **OWL!** (**OWL!**), and **SHACL!** (**SHACL!**) standards to ensure metadata interoperability. A global search function ensures that searching across all available levels of metadata becomes possible.

Archiving The research data and metadata stored in resource types of **RDS!** or Linked Data can be archived for 10 years according to good scientific practice.

2.1 Coscine & FAIR Principles

To enable the reuse of research data in line with FAIR principles across institutional borders, Coscine can be accessed either through participating universities or at a low-threshold level via **ORCID!** [**Haak.2012**]. After registration, researchers can create a research project for which both research data and metadata at various levels are collected and automatically linked. The first level of metadata relates to the research project (including name, description, PIs, discipline). The W3C standards **RDF!** [**Cyganiak.2014**] and **SHACL!** [**Knublauch.2017**] are used for the technical representation and validation of all metadata stored in Coscine. This largely complies with the FAIR principles regarding interoperability and reusability of metadata. In addition, during the life of a project and after its completion, all associated metadata can be publicly shared within Coscine and are searchable and findable. A connection to the NFDI4Ing metadata hub is currently realized via "FAIR Digital Object" interfaces.

In the next step, different data sources, called resources, can be assigned to the research project. For each resource, Coscine assigns a handle-based ePIC-**PID!** [**Kalman.2012**, **Kramer.2016**]. This is used to uniquely and permanently identify the location of the resource and all contained files on a global level. Within resources, fragment identifiers are used to adress individual files. Thus, the research data is permanently referencable and findable in the sense of the FAIR principles.

To date, Coscine has storage resources and Linked Data resources. The storage resources allow researchers to access the **RDS!**, a consortial object storage system funded by the **MKW!** (**MKW!**) and the **DFG!** (**DFG!**). To support researchers' processes as much as possible, Coscine provides multiple ways to interact with research data, either via a browser, using a REST-API or directly via an **S3!** interface. This allows for high performance transfer of even large amounts of research data. When using **RDS!** resources, a retention and archiving period of research data of ten years after the end of

a research project is ensured in terms of **GSP!** and reusability. Within Linked Data resources, externally stored research data is assigned a **PID!** (**PID!**) and can be linked and tagged with metadata. Thus, even for externally stored research data, Coscine allows increasing FAIRness by linking the data with metadata and assigning **PID!**s.

After specifying high-level metadata for the respective resource (including resource name, discipline, keywords, metadata visibility, license), researchers select a suitable selection of metadata fields for their files from various so-called application profiles, e.g. for engineering research data the established EngMeta profile can be used. If a suitable application profile has not yet been added to Coscine, the AIMS Application Profile Generator [**Gronewald.2022**] can be used to create a profile with individual and discipline-specific metadata. Within a resource, researchers can upload their files or store the link to their research data. When using Coscine via the web frontend, file upload is only possible after entering the associated metadata in the application profile. In this way, Coscine makes metadata entry a direct part of the researcher's workflow, supporting the FAIR principles.

Coscine also ensures that data objects and associated metadata, linked by **PID!**, are independently findable and accessible via a REST API. The REST API allows researchers to easily enter their data and metadata into the system and facilitates subsequent use of the same. In addition, the REST API enables token-based authentication to automate workflows. To help researchers interact with Coscine through the interfaces and improve integration with existing data management processes, a team of data stewards and developers has been established to provide tools, programs, and consultation for the technical adaptation of the platform. This includes the collection or extraction of metadata based on the data or the environment in which it was generated. Although the possibilities for automation are highly dependent on the research project, examples and tools support researchers in the implementation and thereby improve the quality of the collected metadata.

Thanks to the interfaces for automation, high technical security standards as well as extensive collaboration possibilities, Coscine is a strong partner for researchers in their daily management of research data. Coscine enables compliance with the FAIR principles from the very first storage of data by bundling raw data, metadata, interfaces and **PID!**s into a linked record according to the "FAIR Digital Objects" model. In this way, Coscine is a valuable contribution to the goal of NFDI4Ing: foster proper research data management in engineering sciences that implements the FAIR data principles.

These layers in Coscine that increase the FAIRness of the research data can be best described with the framework of **FDO!**s (**FDO!**s).

2.2 Coscine & FAIR Digital Objects

The FAIR principles are about making data findable, accessible, interoperable and reusable both for humans and machines. To reach these aims, **RDM!** software requires a framework to store and disseminate digital objects in a robust and informative way. The **FDOF!** (**FDOF!**) provides such a framework by binding all critical information

Figure 1: Research data life cycle

about a digital object: "When a digital object (bit sequence) is identified by a globally unique, persistent and resolvable identifier, characterised by the **FDOF!** typing system and described by metadata records, we can say that we have a FAIR Digital Object." [**Bonino.2022a**]. In this way, **FDO!**s create a new kind of actionable, meaningful and technology independent object that pervades every aspect of life today (https://fairdo.org/).

Although the concept of **DO!** (**DO!**) was introduced by Robert Kahn in the early 1990s, an ecosystem of easy tools that add the **FDO!** layers to raw data including unique identifiers and metadata is still needed (Is FAIR FAIR? An Overview of FAIR Digital Objects – Christine Kirkpatrick, International Data Week 2022, 23.06.2022, https://fairdo.org/library/).

3 How to deal with existing research data / possibilities for automation

Many approaches to research data management consider an ideal scenario where the researchers start from scratch with a new research project. However, this is often not the case, research projects have a very long lifetime and sometimes a correct management of the data and corresponding metadata was not originally considered. Supporting this kind of projects is important as it allows an easier adaption of the research data management platform on a larger scale.

Figure 2: Resources and Metadata linked together

This come with some challenges as there is usually much research data available on different file systems that needs to be gathered and stored within the **RDM!** platform. First step is an analysis of the available data and a collection of the metadata that describes it. Based on this first analysis an application profile can be created that contains all necessary information to store, share and reuse the data later.

Then a suitable resource must be created within Coscine, depending on the requirements of the researchers, different resource types are available. The **RDS-S3!** resource type allows an easy interaction with the underlying storage system, and therefore is suitable for projects with large amounts of already existing research data. The data can be migrated to the **RDM!** platform through a variety of programs, e.g., rclone or minio. These tools can directly upload the data to the underlying s3 bucket. For each bucket there are users with different permission one that can write data and one read only users, thereby also allowing easy reuse of the data.

After uploading the data to Coscine the necessary metadata can be added, the usage of suitable default values can make this process easier. The data can be entered through a form on the website, which also supports editing a batch of files at once. While this approach is feasible, there are more convenient options, especially when working with a lot of files. Coscine comes with an extensive API that allows the usage of all functions that are available on the website through scripts. To secure the access a token is required, which can be created on the website. A token belongs to a specific user and allows the usage of all functions that the user could access through the website. During creation each token is assigned a time frame, in which it is valid, the maximum time frame is one year, thereby ensuring regular revision of the access rights. Of course, every token can be revoked at anytime should a token no longer be required or if it was compromised.

The token can be used to interact with the API, which comes with an extensive documentation of all endpoints, parameters and return values. Swagger is used to allow

the exploration and execution of example queries through a website. An option exists to create curl commands for every query that can be used to a create a custom script to upload the metadata. Through the detailed documentation and the possibility to copy snippets with working queries, even users without a background in computer science can easily use the API and automate parts of their workflow.

Often existing research project often already have research data available that can be extracted from the environment or some files that are stored with the research data. With the tools described above it is possible to write a parser that allows adding the locally available metadata to the files that were uploaded to Coscine.

4 On-boarding of users / Coscine Technical Adaptation

For many researchers Research data management is a new topic. However, the correct handling of metadata and the definition of application profiles is a process that needs experience and an in-depths understanding of the research process and the data. This makes the initial adaptation of an **RDM!** platform difficult for researchers since a certain level of expertise is required to correctly set up a project and the corresponding resources. Usually the most challenging task are the creation of an application profile and applying for storage space.

Tools for the creation of application profiles were created where researchers can use a website to create a new application profile from scratch or explore and extend already existing profiles. If a new profile is created, it will be reviewed by **RDM!** experts thereby ensure a certain quality. Afterwards researchers can apply for storage space. Here an implantation of JARDS (Joint Application Review and Dispatch Service) [**Janetzko.2019**] is used that was adjusted for storage space application. The platform allows researchers to create applications for storage space, these will be reviewed by the maintainers of the storage systems. The review process consists out of multiple stages. At first a formal review is conducted to ensure the application was filled out correctly, then a technical review is done to ensure the feasibility of the application. In case large amounts of storage are requested a scientific review can be performed to ensure the scientific value of the project. JARDS is already widely used within the High-Performance Computing community in Germany, therefore many researchers are already familiar with the platform.

To support researchers with the technical adaptation of the research data management platform Coscine, a group of developers and data Stuarts was created. The group is in direct contact with research groups from different fields and aims at firstly understanding the researchers workflows and process to then suggest new features and improvements. Of course not every workflow can be generalized, however frequent exchange with the researchers allows a better understanding of the requirements and challenges for the adaptation of Coscine.

This groups analyses the needs and workflows of the researchers and provides scripts, programs, tools, and best practices for the interaction with the platform. The material

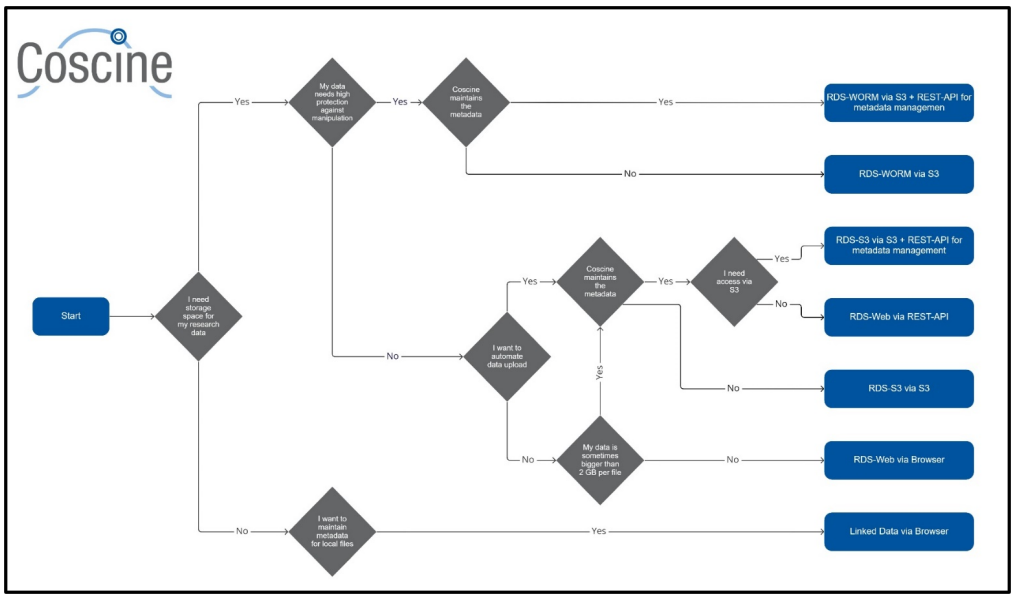

Figure 3: Types of data applications

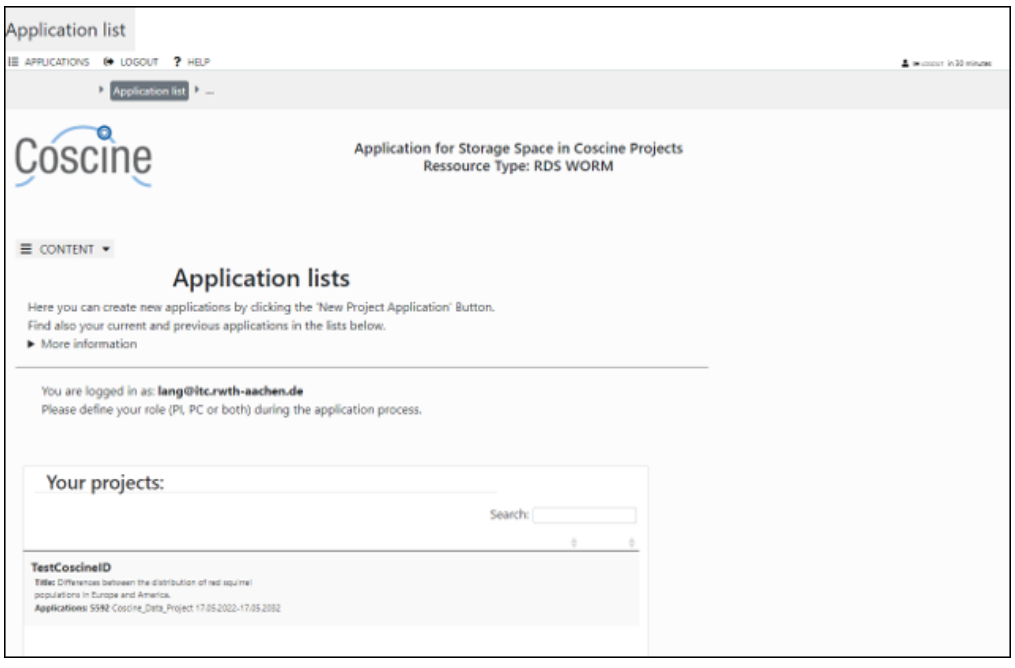

Figure 4: JARDS: Application Overview

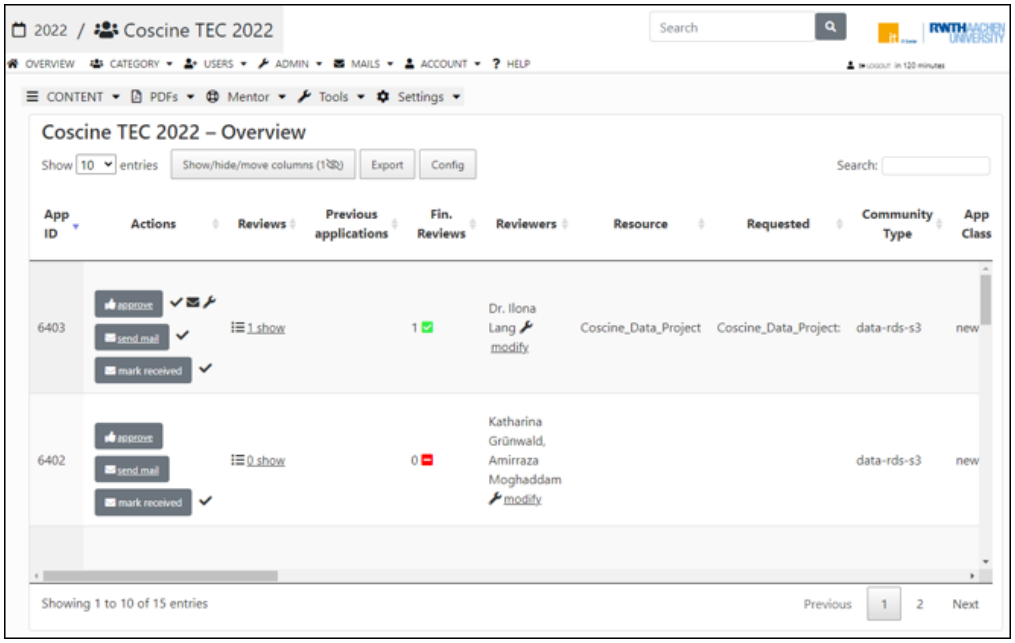

Figure 5: JARDS: Review Overview

is publicly available under an open source license and researchers are encouraged to get involved with the development.

5 Conclusion

Lorem ipsum...

5.1 Limitations

Very specific service provider for single communities Specialized projects that target specific workflows of researchers have a wide acceptance in, often narrow, communities as shown by the different platforms created in TR CRC 32 for geographical data [**Curdt.2014**], medical study data [**Kirsten.2017**] or chemical samples [**Politze.2020**].

Not the whole data life cycle covered so far

5.2 Outlook

This document is an example, two items are cited: *The LaTeX Companion* book: [**latexcompanion**]. And Einstein's journal paper: [**einstein**].

Vivamus vestibulum lacinia laoreet. Pellentesque eu porta massa, a posuere odio. Praesent dolor risus, porta ac ornare ut, lacinia quis est. [1]

6 Acknowledgements

Funding, Funding ([**dirac**]).

1. This is a footnote

7 Roles and contributions

Marcel Nellesen: Conceptualization, Writing – original draft

Marius Politze: Conceptualization, Supervision, Project administration