1 Introduction

Universities and enterprises worldwide perform research to improve the current state of the art. This research produces a lot of diverse data called research data. Since research results are usually based on such data, the importance of having guidelines for research data becomes apparent. For this reason, the FAIR Guiding Principles [1] were defined. They are the foundational principles that research data management should adhere to. In summary, they define that data should be:

- Findable
  The data can be easily found when looking for it.
- Accessible
  There is a simple way to access data: publicly or with authentication.
- Interoperable
  Information about the data is accessible and provided in a machine-readable and standardized way.
- Reusable
  The data is reusable for future work, and its metadata is thoroughly described.

Research data management (RDM) teams are established at universities to facilitate these FAIR principles and build solutions to simplify research processes [2–4]. Thereby, new ways of conducting processes are introduced in faculties, making them face new challenges and keeping this a very active research area.

1.1 Motivation

Because research data management (RDM) is a recent and essential topic, one significant task is to get an overview of the research data being produced daily. Whenever researchers try to access research data from someone else, there should be no uncertainty about the data's whereabouts. Furthermore, there is a need for reproducibility because researchers might want to replicate the findings of a paper. Therefore, they expect an easy way to determine how the research data came together. For these challenges, one idea is to incorporate the FAIR principles [1]. However, a primary challenge in RDM is that research data moves through the research data life cycle, as shown in Figure 1.1 and described by [5]. It starts with the planning process and ends with reusability concerns. This means that research data goes through different stages, adaptations, and even locations during these phases, increasing the complexity of fulfilling the FAIR principles. Furthermore, changes in research data are hard to follow because they are usually done manually by researchers without any platform being able to take notice of them.

Previous approaches at institutions like the RWTH Aachen University have been individual systems that existed for single phases or parts, like simpleArchive for *Storage* [7], Sciebo for *Production* [8], and the Metadata Manager for metadata management [9]. However, with no connection between them, it became apparent that movement cannot be tracked, and adhering to the FAIR principles can only happen on the individual stage. For this reason, more comprehensive solutions that incorporated a larger part of the research data life cycle needed to be created. One of those solutions at the RWTH Aachen University is called Coscine [10]. The main goal of Coscine is to incorporate multiple phases by being a platform that can incorporate any number of storage providers. It essentially combines the efforts of [7], [8], and [9] and brings them together as one comprehensive product. Thereby, it enables tackling the FAIR principles across multiple stages. It should be

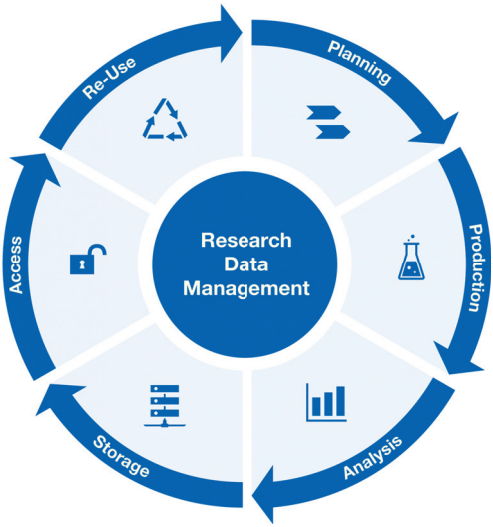

Figure 1.1: Research data life cycle from [6].

noted that other institutions also try to incorporate multiple stages in one application, like the KIT Data Manager [11], the RADAR system [12], or the MASi repository [13]. Additionally, electronic lab notebooks (ELNs) like eLabFTW [14] and Chemotion [15] have been created, which try this as well.

Data provenance is one of the most important aspects of facilitating the FAIR principles. It means that the path data has traversed should be recorded, the changes noted, and the location defined. For tracking research data, one part is to attach a persistent identifier (PID) and update it with every movement. One such system that offers PIDs is ePIC [16]. However, the changes still need to be described, especially regarding the Five Ws (who, what, when, where, and why). Doing this manually with a provenance ontology like PROV-O [17] is usually fine if all traffic goes through the platform that provides and manages data. However, for platforms like Coscine, the research data might be stored with a storage provider (like an S3 bucket), where data access is mostly not going through

the platform itself but directly through the storage provider. This makes describing the changes in research data difficult because many changes will never be noticed. Therefore, this thesis identifies this issue and proposes the need to determine provenance after the change has happened. In the following, the issue will be called asynchronous data provenance. This thesis's main goal is to establish asynchronous data provenance in an environment where no data provenance can be found. This requires certain building blocks to compare different versions of research data. That is why a representation of research data, like extracted interoperable metadata, is necessary. With this, changes between versions should be accurately identified. This work will be integrated into a use case (Coscine) that acts as a research data management system and is adapted during this thesis. These efforts are being taken to improve the quality of research data and come closer to fulfilling the FAIR principles in an environment where this is a difficult challenge.

1.2 Related Research

To fully present an overview of this thesis, the current state of the art needs to be described. The thesis's topic is placed in five research areas and identifies gaps it aims to close. The relevant areas are FAIR Digital Objects, data provenance, metadata extraction, similarity detection, and distributed research data management systems.

1.2.1 FAIR Digital Objects

At the center of this whole thesis is research data. For this, it is necessary to define what research data is and what the specific parts of it are. *Research data* is an over-encompassing term that can mean multiple datasets or a single item produced in some research processes, depending on the context. It is by that very

general. When talking about parts of data, the literature provides a couple of terms like data element [18], digital entity [19], or digital object [20]. However, a new term has been coined recently, the FAIR Digital Object (FDO) [21], as following the FAIR principles in research data management is essential. An FDO is a digital object that fully fulfills the FAIR principles. An illustration of an FDO is shown in Figure 1.2.

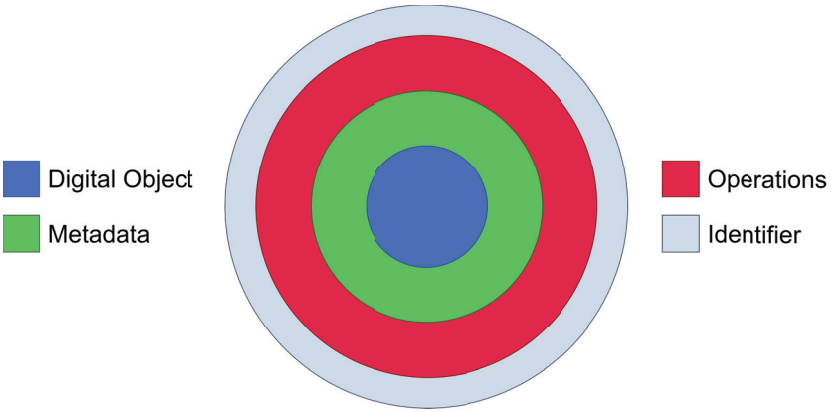

Figure 1.2: Illustration of an FDO based on [21].

The illustration in Figure 1.2 details the necessary encapsulations a digital object needs to have to fulfill the FAIR principles. The capsules are metadata records that describe the digital object, operations that provide (standardized) interaction points with a digital object, and a persistent identifier that points to the digital object.

The FDO topic has recently gained much attention with the Leiden Declaration on FAIR Digital Objects [22] and the 1st International Conference on FAIR Digital Objects in 2022 [23]. Research is currently being performed to comply with the goals of FDOs and enhance the current state [24].