TU WIEN Informatics

Erweiterung von Datenrepositorien mit dem modernen FAIR Digital Objects Framework

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Software Engineering & Internet Computing

eingereicht von

Bartosz Błachut, BSc.
Matrikelnummer 12249629

an der Fakultät für Informatik
der Technischen Universität Wien
Betreuung: Dr.techn. Mag. Tomasz Miksa

Wien, 20. April 2025

Bartosz Błachut	Tomasz Miksa

TU WIEN Informatics

Extending data repositories with a novel FAIR Digital Objects framework

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Software Engineering & Internet Computing

by

Bartosz Błachut, BSc.
Registration Number 12249629

to the Faculty of Informatics

at the TU Wien

Advisor: Dr.techn. Mag. Tomasz Miksa

Vienna, April 20, 2025

Bartosz Błachut	Tomasz Miksa

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Erklärung zur Verfassung der Arbeit

Bartosz Błachut, BSc.

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.
Ich erkläre weiters, dass ich mich generativer KI-Tools lediglich als Hilfsmittel bedient habe und in der vorliegenden Arbeit mein gestalterischer Einfluss überwiegt. Im Anhang „Übersicht verwendeter Hilfsmittel“ habe ich alle generativen KI-Tools gelistet, die verwendet wurden, und angegeben, wo und wie sie verwendet wurden. Für Textpassagen, die ohne substantielle Änderungen übernommen wurden, habe ich jeweils die von mir formulierten Eingaben (Prompts) und die verwendete IT-Anwendung mit ihrem Produktnamen und Versionsnummer/Datum angegeben.

Wien, 20. April 2025

Bartosz Błachut

Danksagung

Hiermit würde ich gerne mich bei meinem Arbeitsbetreuer herzlich bedanken—Tomasz Miksa. Ich bin extrem glücklich, dass ich Ihnen an der TU Wien begegnet habe.

Acknowledgements

Hereby, I would like to express my gratitude towards my thesis advisor—Tomasz Miksa.
I am extremely grateful that I had a chance to meet you on the TU Wien.

Kurzfassung

Die FAIR-Prinzipien—Findable, Accessible, Interoperable und Reusable—haben sich als weithin anerkannte Richtlinien zur Verbesserung des Forschungsdatenmanagements etabliert. Ihr Ziel ist es, insbesondere im Kontext von Metadaten, die maschinelle Verarbeitbarkeit und semantische Klarheit zu fördern. Obwohl verschiedene Metriken und Werkzeuge zur Bewertung der FAIRness von Datensätzen existieren und zahlreiche Initiativen zur Umsetzung FAIR-konformer Lösungen entstanden sind, bleiben die Definitionen breit und abstrakt. Dies hat zu vielfältigen Interpretationen und Implementierungen in Datenrepositorien geführt.

Ein vielversprechender Ansatz ist das Konzept der FAIR Digital Objects (FDOs), das vom FDO Forum als strukturierte Datenobjekte mit FAIR-konformen beschreibenden Elementen eingeführt wurde. Die FDO-Spezifikation legt fest, wie solche Objekte aufgebaut sein sollen, wurde jedoch bislang in keinem repräsentativen Forschungsdatenrepositorium als Teil der Metadateninfrastruktur umgesetzt. Zudem wurde die FAIRness von Datensätzen, die auf diese Weise modelliert wurden, bisher nicht empirisch untersucht. Wir stellen die Hypothese auf, dass die Implementierung einer auf der FDO-Spezifikation basierenden Repositoriumserweiterung die FAIRness-Bewertung von Metadatensätzen im Vergleich zu bereits existierenden Ansätzen verbessern kann.

In dieser Arbeit präsentieren wir eine neuartige FDO-Infrastruktur, die auf der Cordra-Plattform gehostet wird und speziell zur Repräsentation von Metadaten aus Forschungsdatenrepositorien konzipiert ist. Diese Infrastruktur umfasst ein wiederverwendbares Kern-Framework sowie eine *Research Metadata Extension* basierend auf dem DataCite-Metadatenschema. Wir integrieren diese Lösung in InvenioRDM mithilfe eines eigens entwickelten Werkzeugs—dem *Migration Assistant*—und bewerten die FAIRness der migrierten Datensätze unter Verwendung der FAIRsFAIR-Metriken und des FUJI-Bewertungswerkzeugs. Unsere Ergebnisse zeigen eine messbare Steigerung der FAIRness, insbesondere in den Bereichen Maschinenlesbarkeit und semantische Interoperabilität, und liefern ein konkretes, reproduzierbares Modell zur FAIR-konformen Metadatenrepräsentation.

Abstract

The FAIR principles—Findable, Accessible, Interoperable, and Reusable—have become a widely recognized set of guidelines for improving research data management. They aim to enhance machine-actionability and semantic clarity, especially in the context of metadata. Although various metrics and tools exist to evaluate the FAIRness of data records, and many initiatives have emerged to implement FAIR-aligned solutions, the definitions remain broad and abstract. This has led to diverse interpretations and implementations across data repositories.

One promising direction is the concept of FAIR Digital Objects (FDOs), introduced by the FDO Forum as structured data entities that are equipped with FAIR descriptive elements. The FDO Specification outlines how such objects should be constructed, yet no representative research data repository has implemented them as part of its metadata infrastructure. Furthermore, the FAIRness of records modeled in this way has not been empirically assessed. We hypothesize that implementing a repository extension based on the FDO Specification can increase the FAIRness score of metadata records compared to already existing approaches.

In this thesis, we present a novel FDO infrastructure hosted on the Cordra platform and tailored to represent metadata from research data repositories. This infrastructure includes a reusable core framework and a *Research Metadata Extension* based on the DataCite Metadata Schema. We integrate it with InvenioRDM through a custom-built tool—the *Migration Assistant*—and evaluate the FAIRness of migrated records using FAIRsFAIR metrics and the FUJI assessment tool. Our results show a measurable increase in FAIRness, particularly in aspects of machine-readability and semantic interoperability, providing a concrete and reproducible model for FAIR-compliant metadata representation.

Contents

CHAPTER 1

Introduction

Numerous institutions and organizations such as the Research Data Alliance (RDA)[1] or Committee on Data of the International Science Council (CODATA)[2] aim to improve *the ability of machines to automatically find and use* [WDA+16] scholarly data. One of the important developments in this area recognized by the mentioned institutions are the FAIR principles introduced in 2016 by Wilkinson et al. in [WDA+16]. The name FAIR is an acronym and stands for findable, accessible, interoperable and reusable—four core terms that represent the proposed guidelines. Wilkinson et. al claim that following of these guidelines is supposed to contribute to enhancing data reusability and, most importantly, its machine-actionability.

Researchers publish their findings and reports—their scholarly data—in dedicated repositories, such as the Dryad Digital Repository[3], the Open Science Framework[4], Zenodo[5] or the TU Wien Research Data (TUWRD)[6]—we refer to these as research data repositories. Since the FAIR principles have been widely acclaimed and embraced by the Open Science movement and the European Union [Gil24, HJC+18], it has become a standard among researchers and institutions worldwide to publish scholarly data following these guidelines. Therefore, when publishing research data in a repository, it is important to provide a description—metadata—in a FAIR way. However, standardizing such a FAIR metadata format might prove to be difficult since Wilkinson et al. do not propose a set of strict rules but rather inspiring concepts that can be interpreted and implemented on a multitude of ways [Gro20].

[1] https://www.rd-alliance.org/
[2] https://codata.org/
[3] https://datadryad.org/
[4] https://osf.io/
[5] https://zenodo.org
[6] https://test.researchdata.tuwien.ac.at/

Notwithstanding, there are initiatives that aim to assess the so called FAIRness of a certain record from a research data repository—how FAIR this record is. One of such initiatives is the FAIRsFAIR project, in terms of which a set of metrics that can be used for such a FAIRness evaluation has been published. There also are projects that attempt to standardize FAIR implementations in form of recommendations.

One of such recommendations are the FAIR Digital Objects (FDOs)—a concept specified by the FDO Forum in [ICD+23]. It is a notion of structured data entities—Digital Objects—that are equipped with certain descriptive attributes. The structure of these attributes has to strictly follow certain guidelines introduced in the FDO Specification [ICD+23]. This, in turn, is supposed to make these objects understandable and interpretable to machines—it is supposed to make them FAIR. The definition of FDOs is based around the FAIR principles, which makes the whole idea a notable FAIR recommendation. However, the FDO Specification is purely theoretical and provides no concrete implementation recommendations. Moreover, there are no representative research data repositories that would represent their records' metadata as FAIR Digital Objects and the FAIRness of such a solution has not yet been measured.

Additionally, the FAIRness of records published in the aforementioned representative research data repositories varies. Using a tool that establishes the FAIRness as a score on a percentage scale (from 1 to 100), we have conducted certain experiments and established that the FAIRness of most of the published records is below 80%. We have presented the said experiments in Section 2.3.

In this thesis, we attempt to increase the FAIRness level of records in a research data repository by representing their metadata as FAIR Digital Objects. For this purpose, we develop a novel FDO infrastructure with a set of descriptive entities that allow us to appropriately model records' metadata. It is our concrete interpretation of the FDO Specification and it is designed to be hosted on the Cordra open source Digital Object management software[7]. This infrastructure contains two elements: a core FDO framework that can be fine-tuned and reused for other purposes and an extension of this framework that allows us to model the metadata of records from a research data repository—the *Research Metadata Extension*. We have designed the latter part of our proposed FDO infrastructure with focus on metadata that follows a widely adopted standard—the DataCite Metadata Schema[8][LRA+24].

Furthermore, we propose a conceptual architecture of a research data repository that represents its records' metadata as FDOs using the aforementioned novel FDO infrastructure. We also provide a migration procedure that can be used to migrate the metadata from an already existing research data repository to our FDO infrastructure.

To assess the effectiveness of our solution, we create a tool that allows us to extend an already existing research data repository and represent its records' metadata as FDOs.

[7] https://www.cordra.org/
[8] https://schema.datacite.org/

We focus on repositories based on the InvenioRDM[9]—a popular research data repository framework that follows the DataCite Metadata Schema. We evaluate the proposed novel metadata representation against the FAIRsFAIR metrics using the data from the test instance of TU Wien Research Data (TUWRD). Test results confirm that the proposed solution increases the FAIRness score of the assessed records.

Altogether, this thesis presents the following contributions.

1. A novel FDO framework that has been designed by strictly following the FDO Specification [CML+25]. It is important to note that this framework can also be used for other use cases than the one studied in this thesis. It consists of 6 core FDOs and it is designed to be hosted on the Cordra open source Digital Object management software.

2. An extension of the introduced FDO framework—a set of FDOs that are crucial to describe the metadata of records from a research data repository. Together with the said FDO framework, this extension is a concrete interpretation of the FDO Specification for a given purpose—to model the metadata of records that follow a widely adopted standard—the DataCite Metadata Schema.

3. A conceptual architecture of a research data repository that represents its records' metadata as FDOs using the aforementioned novel FDO infrastructure. We also provide a migration procedure that can be used to migrate the metadata from an already existing research data repository to our FDO infrastructure.

4. Our recommendations for possible improvements of data repository software and FAIR testing tools, especially to the InvenioRDM community and users of the FAIRness assessment tool that we used—FUJI. These recommendations are a result of evaluation of our solution using the data from the test instance of TUWRD.

1.1 Expected results

The expected results of this master thesis are formulated as research questions that are presented in the following sections.

1.1.1 What are the appropriate FDO profiles and attributes that can be used to represent records in a research data repository as FDOs?

An important part of this thesis is to design appropriate parts of an FDO system—such as object attributes, attribute definitions and profiles—strictly following the FDO specification [CML+25] maintained by the FDO Forum[10]. Another crucial aspect is to choose

[9] https://inveniosoftware.org/products/rdm/
[10] https://fairdo.org/

and design FDO types as well as Linked Open Data (LOD)[11] vocabularies to correctly model the relationships between the existing entities in the system.

We have split this effort into two parts. Firstly, we designed an abstract FDO framework that consists of 6 core building blocks. This framework can be adjusted and reused for other purposes. We expand upon how it is structured in Section 3.2. Secondly, we designed an extension on top of this framework—the *Research Metadata Extension* that allows us to model research data repository records' metadata as FDOs. Our work regarding the extension has been presented in Section 3.3.

Together, these two elements—the FDO framework and the *Research Metadata Extension*—are the novel FDO infrastructure that we created for the purpose of this thesis. It is designed to be hosted on the Cordra open source Digital Object management software and it can be used to model metadata that follows the DataCite Metadata Schema.

1.1.2 What is the conceptual architecture of a research data repository based on InvenioRDM that represents datasets as FDOs?

In this thesis, we present the conceptual architecture of a research data repository that represents its records as FAIR Digital Objects. We propose that it should contain two services—a research data repository and an FDO backend for storing the FAIR Digital Objects. The two services should be connected to each other using a third service that serves as a translation layer between them. We have expanded upon this in Section 3.1.

To evaluate the effectiveness of our conceptualization, we design and implement a concrete example. We choose the Cordra open source Digital Object management software as the backend for the FDOs and we focus on repositories based on a popular research data repository framework—the InvenioRDM project. We create an application—the *Migration Assistant*—that serves as a translation layer between the two chosen services. This tool allows for a straightforward migration of the metadata of the records from a chosen repository based on the InvenioRDM project to the Cordra backend, where the metadata is represented as FAIR Digital Objects. It leverages the appropriate REST APIs of the two services and it exposes an additional endpoint for evaluation of the FAIRness of the new metadata format. This endpoint uses the Digital Object Interface Protocol (DOIP) Cordra API, which means that the newly created FAIR Digital Objects are being accessed following the rules of the Digital Object Architecture (DOA).

The *Migration Assistant* tool together with the two services—a research data repository based on InvenioRDM and the Cordra backend—are the architecture of a research data repository that represents its records' metadata as FDOs. This architecture has been presented in Section 4.1.

[11] We explain the concept of Linked Open Data in Section 2.1

1.1.3 What are the advantages of adopting the proposed solution with regard to the FAIR principles?

The solution developed for the purposes of this thesis should be evaluated in terms of satisfying the FDO specification and the impact it has on the FAIRness of the records in an existing repository. We aim to increase this FAIRness both theoretically—by improving the machine-actionability of the data in the repository—and practically—we attempt to show that applying our solution alongside an already existing repository increases the FAIRness of its records with regard to the FAIRsFAIR metrics.

First of all, all FDO specification requirements are met, so extending an existing repository with the provided solution is in itself an advantage, because the repository records will become machine-actionable in a way that follows the newest industry standards. This elevates the overall FAIRness status of the data in the extended repository.

Second of all, the solution does not decrease the FAIRness level of the records in target data repositories. On the contrary, a significant increase of this level has been observed in the evaluation of our solution. FAIRsFAIR metrics have been applied to evaluate how the application of the proposed solution influences the FAIRness scores of previously existing records. We have tested 210 records from the test instance of TU Wien Research Data using FUJI—an automated tool that assesses a record's FAIRness with regard to the FAIRsFAIR metrics—and we established that the FAIRness score indeed increases after the application of the proposed solution. In total, every record from the TU Wien Research Data repository achieved a higher score with the proposed metadata format compared to metadata extracted directly from TUWRD. We provide a detailed report regarding these findings in Section 4.2 and Section 4.3.

1.2 Methodology

While working on this thesis, the author followed an approach described in the following steps:

1. **Literature research**
   A significant amount of effort had to be put into meticulous background research. This step is fundamental to establish the premises of other FDO implementation recommendations, such as the ones proposed by the WorldFAIR Foundation. It is also important to choose the appropriate FDO schemas and LOD technologies that will be used to correctly represent metadata of the records in a research data repository. Additionally, the author needed to gain significant amount of knowledge about the InvenioRDM and Cordra projects.

2. **Design**
   The knowledge gathered in the previous stage had to be assembled together to create the desired framework. Further design decisions regarding attributes, attribute definitions, profiles and FDO schemas had to be made based on the selected set of

records published in the TU Wien Research Data. Moreover, the architecture of the solution—how the InvenioRDM and Cordra components would be connected to each other—had to be conceptualized and visualized.

3. **Implementation**
   In this step, the design of the entire framework had to be materialized in form of a working implementation. Furthermore, exemplary research data records should be added to the created system to allow for the evaluation measures from the next step.

4. **Evaluation**
   The prepared solution had to be evaluated against the FDO specification and the FAIR principles. First of all, it had to be established if the solution followed the core guidelines of the FDO specification. Secondly, FAIRsFAIR metrics had to be applied to measure the impact this solution had on the FAIRness level of previously existing repository records.

1.3 Thesis structure

This thesis is structured in the following five chapters and two appendices.

1. **Chapter 1**—Introduction. This is an introductory chapter that is an explanation of the purpose of our work and we talk about the motivation behind this thesis. We also provide three research questions that represent the expected results of this project. Furthermore, we also mention how our work is structured and what is the used methodology.

2. **Chapter 2**—Related work. In the first chapter, we introduced some concepts that are closely related to our work. This chapter is a more detailed background about these technologies. We explain in more depth what stands behind FAIRsFAIR, FUJI, FAIR Digital Objects (FDOs), Cordra and the InvenioRDM project.

3. **Chapter 3**—Proposed Solution. In this chapter, we describe the details of:

   - the novel FDO framework,
   - the FDOs that are necessary to describe the metadata of records from a research data repository,
   - the conceptualization of a procedure that migrates the metadata of records from a research data repository to an FDO backend.

   Moreover, we provide answers to the first research question from the previous chapter.

4. **Chapter 4**—Evaluation. In this chapter, we present our application that migrates the records' metadata from an existing InvenioRDM instance to the Cordra backend

and we provide details about how we assessed the initial FAIRness of the records from InvenioRDM and what elements of the metadata should be altered to improve the score. We also evaluate how the FAIRness score changes after introducing our solution and describe in detail the factors that directly contribute to this change. Moreover, we provide answers to the second and third research questions from the previous chapter.

5. **Chapter 5**—Conclusion. This chapter is a succinct summary of this entire thesis.

6. **Appendix A**—Source Code. In this appendix we include references to the implementation of our application developed for the purpose of this thesis.

7. **Appendix B**—Test Results. In this appendix, we reference a repository with the results of evaluation carried out using the FUJI tool.

CHAPTER 2

Related work

This chapter explains in more detail the concepts and technologies that are closely related to our work. We provide an in-depth background about the FAIR principles themselves and focus on the FAIR recommendations and projects that are relevant to this thesis. We expand upon the concept of FAIRness and introduce information about the FAIRness indicators, such as the ones presented in the FAIRsFAIR project. Moreover, we also discuss the FAIR Digital Objects and provide more insight about this particular FAIR recommendation. Furthermore, we present an overview of the technologies that we analyzed and used in this thesis—InvenioRDM and Cordra.

2.1 FAIR

Since its inception in 1994, the World Wide Web Consortium (W3C)[1] aims to develop standards and guidelines to extend the World Wide Web. One of the key efforts of the W3C is publishing recommendations regarding the Semantic Web[2]—a concept introduced in 2001 by Tim Berners-Lee [BLHL01], who initially envisioned it as a web of data *in which information is given well-defined meaning, better enabling computers and people to work in cooperation.* Berners-Lee's main goals, thus, were to model data provenance and semantics—the metadata—in a way that is understandable not only to humans but also machines. To facilitate for these objectives, technologies, which are today known under the umbrella term: Linked Open Data (LOD)[3] [BHBL09, BL06], such as the Resource Description Framework (RDF) [CLW14], JSON-LD format [SKL14] and the Web Ontology Language (OWL) [PSMP12], have been introduced.

The Semantic Web offers, most importantly, interoperability of data and its metadata. Information is not published alone, but accompanied by a *distributable, machine-readable*

[1] https://www.w3.org/
[2] https://www.w3.org/2001/sw/
[3] https://www.w3.org/2013/data/

description, which is a primary building block of the distributed web of data that we are able to witness today [AHG20]. The existing LOD technologies address problems related to integrating different data sources by introducing shared semantics between multiple services and standardizing the used terminology in publicly available vocabularies [QBC13].

The notion of machine readability and findability, which is offered by the Semantic Web, is considered desirable by a variety of institutions and organizations such as the Research Data Alliance (RDA)[4] or Committee on Data of the International Science Council (CODATA)[5], which aim to improve findability and understandability of scholarly data. One of the important developments recognized by the mentioned institutions are the FAIR principles introduced in 2016 by Wilkinson et al. in [WDA+16]. The name—FAIR—is an acronym which stands for: findable, accessible, interoperable and reusable—the core guidelines which are also divided into more granular sub-concepts. Wilkinson et al. claim that following of these guidelines is supposed to contribute to enhancing data reusability and, most importantly, *the ability of machines to automatically find and use data.*

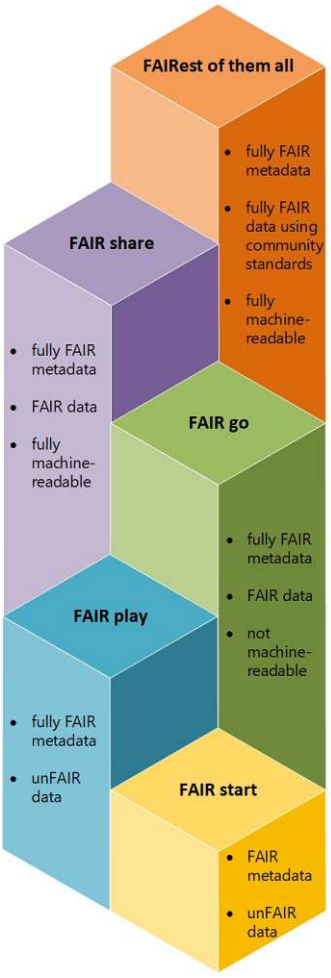

Figure 2.1: Five maturity levels introduced by JRC in [HTL+25].

The publication of the FAIR guidelines has sparked new research in this direction. Many recommendations and tools have been created to foster adoption of these principles by institutions and researchers. An important development regarding the FAIR principles are the FAIR Digital Objects (FDOs) [ICD+23]. FDOs are a conceptualization of how a Digital Objects framework should be structured for it to become FAIR. FDOs represent data entities that are equipped with FAIR descriptive elements. Their definition is based around the FAIR principles, which makes the whole idea a notable proposal on how a FAIR system should be constructed.

Important to note are also the FAIR indicators published by RDA in [Gro20]. In this document, a set of guidelines for assessing the extent to which the FAIR principles are satisfied has been specified. These indicators have been designed for re-use by others in their evaluation approaches. They have been adopted in projects sponsored by the European Union. For instance, the Joint Research Center (JRC), to increase the FAIRness of their data,

[4] https://www.rd-alliance.org/
[5] https://codata.org/

have introduced five FAIRness maturity levels in [HTL+25] that have been depicted in Figure 2.1. Each level is accompanied by a detailed description and examples.

There also are other projects that aim to propose a standardized FAIR framework which could be adopted by a significant amount of researchers. The WorldFAIR[6] Foundation (introduced by Committee on Data of the International Science Council (CODATA) and founded by the European Union [Mol22]) provides recommendations on standardizing FAIR implementations to support all areas of research. Their core project—Cross-Domain Interoperability Framework (CDIF) [GBB+24] is based on use cases from different scientific fields and aims to create a lingua franca among many FAIR implementations. However, this remains only a set of recommendations and not a concrete design and implementation of a FAIR system. Applying these recommendations to an already existing repository system might require a significant amount of manual labor.

Given the fact that increasing the interoperability and machine-actionability of data is a broad and open topic, it is difficult to quantify if other similar solutions exist. Moreover, approaches beyond the FAIR principles exist as well. The so-called Research Object Crate (RO-Crate)[7] [ea22] is one of such efforts maintained by a broad community of researchers. RO-Crate's specification is based around a concept of a *Research Object*—a structure similar to a Digital Object—and concretely standardizes its metadata in a way that follows the assumptions of LOD. However, it is not based around the FAIR principles and there is no guarantee that conforming to this standard will increase the FAIRness level of data.

2.2 FAIRsFAIR

Another adaptation of the RDA indicators (mentioned in Section 2.1) has been presented as part of the FAIRsFAIR project. A set of 17 FAIRsFAIR Data Object Assessment Metrics[8] has been published in [DHM+22]. They can be used to evaluate the so called FAIRness level of a given research data record. The metrics with their appropriate descriptions from [DHM+22] have been depicted in Table 2.1.

FAIR principle	FAIRsFAIR metric	description
findability (F)	*FsF-F1-01D*	Data is assigned a globally unique identifier.
	FsF-F1-02D	Data is assigned a persistent identifier.

[6] https://worldfair-project.eu/
[7] https://www.researchobject.org/ro-crate/
[8] https://www.fairsfair.eu/fairsfair-data-object-assessment-metrics-request-comments

	FsF-F2-01M	Metadata includes descriptive core elements (creator, title, data identifier, publisher, publication date, summary and keywords) to support data findability.
	FsF-F3-01M	Metadata includes the identifier of the data it describes.
	FsF-F4-01M	Metadata is offered in such a way that it can be retrieved by machines.
accessibility (A)	*FsF-A1-01M*	Metadata contains access level and access conditions of the data.
	FsF-A1-02M	Metadata is accessible through a standardized communication protocol.
	FsF-A1-03D	Data is accessible through a standardized communication protocol.
	FsF-A2-01M	Metadata remains available, even if the data is no longer available.
interoperability (I)	*FsF-I1-01M*	Metadata is represented using a formal knowledge representation language.
	FsF-I2-01M	Metadata uses semantic resources.
	FsF-I3-01M	Metadata includes links between the data and its related entities.
reusability (R)	*FsF-R1-01MD*	Metadata specifies the content of the data.
	FsF-R1.1-01M	Metadata includes license information under which data can be reused.
	FsF-R1.2-01M	Metadata includes provenance information about data creation or generation.
	FsF-R1.3-01M	Metadata follows a standard recommended by the target research community of the data.
	FsF-R1.3-02D	Data is available in a file format recommended by the target research community.

Table 2.1: FAIRsFAIR metrics along with their appropriate descriptions from [DHM+22].

To measure findability (F), the *FsF-F1-01D*, *FsF-F1-02D*, *FsF-F2-01M*, *FsF-F3-01M* and *FsF-F4-01M* metrics should be applied. The first two require that a globally unique and persistent identifier is assigned to the data. The ones ending with *M* regard the metadata and demand that it is retrievable by machines and that it includes a set of chosen descriptive elements and an identifier of the data it describes. Similarly, accessibility (A) is expressed in: *FsF-A1-01M*, *FsF-A1-02M*, *FsF-A1-03D* and *FsF-A2-01M*—the metadata and data should be accessible through a standardized communication protocol and the

metadata has to contain access conditions to the data. Moreover, the metadata should be persisted, even if the data is no longer available. For interoperability (I), *FsF-I1-01M*, *FsF-I2-01M* and *FsF-I3-01M* have been created. In accordance with those metrics, the metadata should be represented by a formal knowledge representation language, use semantic resources and include links between the data and its related entities. Finally, reusability (R) should be measured using: *FsF-R1-01MD*, *FsF-R1.1-01M*, *FsF-R1.2-01M*, *FsF-R1.3-01M* and *FsF-R1.3-02D*. These last few metrics make sure that the metadata specifies content of the data and includes license and provenance information regarding the data it describes. Moreover, metadata should follow a standard recommended by the target research community of the data and the data itself should be in a file format recognized by this community.

To ease the evaluation of an existing implementation against the aforementioned metrics, an automated online tool—FUJI[9] [DH20]—has been introduced. It analyzes a given research data entry from a certain repository and presents a detailed analysis of its FAIRness based on the FAIRsFAIR metrics along with a FAIRness percentage score. An example output of this tool has been depicted in Figure 2.2. It evaluates a given data record against 16 out of 17 FAIRsFAIR metrics—the *FsF-A2-01M* (metadata remains available, even if the data is no longer available) metric has not been included. Such an assessment, however, remains to this day a complicated challenge, given the fact that there exist multiple possible metadata implementations and data infrastructure frameworks.

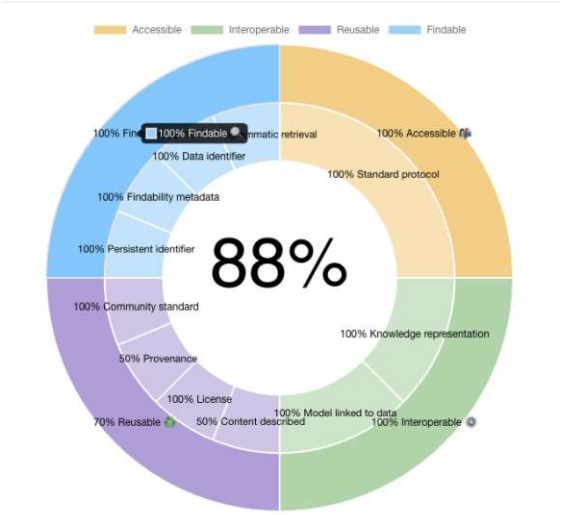

Figure 2.2: Example result of evaluation using the FUJI tool.

To perform its task, FUJI executes 47 tests in total. They have been depicted in Table 2.2 along with their appropriate descriptions.

FAIRsFAIR metric	FUJI test	description
FsF-F1-01D	*FsF-F1-01D-1*	Identifier is resolvable and follows a defined unique identifier syntax (IRI, URL).
	FsF-F1-01D-2	Identifier is not resolvable but follows an UUID or HASH type syntax.
FsF-F1-02D	*FsF-F1-02D-1*	Identifier follows a defined persistent identifier syntax.
	FsF-F1-02D-2	Persistent identifier is resolvable.

[9] https://www.f-uji.net/

FsF-F2-01M	*FsF-F2-01M-1*	Metadata has been made available via common web methods.
	FsF-F2-01M-2	Core data citation metadata is available.
	FsF-F2-01M-3	Core descriptive metadata is available.
FsF-F3-01M	*FsF-F3-01M-1*	Metadata contains data content related information (file name, size, type).
	FsF-F3-01M-2	Metadata contains a PID or URL which indicates the location of the downloadable data content
FsF-F4-01M	*FsF-F4-01M-1*	Metadata is given in a way major search engines can ingest it for their catalogs (embedded JSON-LD, Dublin Core or RDFa).
	FsF-F4-01M-2	Metadata is registered in major research data registries (DataCite).
FsF-A1-01M	*FsF-A1-01M-1*	Information about access restrictions or rights can be identified in metadata.
	FsF-A1-01M-2	Data access information is machine-readable.
	FsF-A1-01M-3	Data access information is indicated by (not machine-readable) standard terms.
FsF-A1-02M	*FsF-A1-02M-1*	Landing page link is based on standardized web communication protocols.
FsF-A1-03D	*FsF-A1-03D-1*	Metadata includes a resolvable link to data based on standardized web communication protocols.
FsF-I1-01M	*FsF-I1-01M-1*	Parsable, structured metadata (JSON-LD, RDFa) is embedded in the landing page XHTML/HTML code.
	FsF-I1-01M-2	Parsable, graph data (RDF, JSON-LD) is accessible through content negotiation, typed links or SPARQL endpoint.
FsF-I2-01M	*FsF-I2-01M-1*	Vocabulary namespace URIs can be identified in metadata.
	FsF-I2-01M-2	Namespaces of known semantic resources can be identified in metadata.
FsF-I3-01M	*FsF-I3-01M-1*	Related resources are explicitly mentioned in metadata.
	FsF-I3-01M-2	Related resources are indicated by machine-readable links or identifiers.
FsF-R1-01MD	*FsF-R1-01MD-1*	Minimal information about available data content is given in metadata.

	FsF-R1-01MD-1a	Resource type (e.g. dataset) is given in metadata.
	FsF-R1-01MD-1b	Information about data content (e.g. links) is given in metadata.
	FsF-R1-01MD-2	Verifiable data descriptors (file info, measured variables or observation types) are specified in metadata.
	FsF-R1-01MD-2a	File size and type information are specified in metadata.
	FsF-R1-01MD-2b	Measured variables or observation types are specified in metadata.
	FsF-R1-01MD-2c	Data service endpoint and protocol information are specified in metadata.
	FsF-R1-01MD-3	Data content matches file type and size or protocol specified in metadata.
	FsF-R1-01MD-4	Data content matches measured variables or observation types specified in metadata.
FsF-R1.1-01M	*FsF-R1.1-01M-1*	License information is given in an appropriate metadata element.
	FsF-R1.1-01M-2	Recognized license is valid (community specific or registered at SPDX).
FsF-R1.2-01M	*FsF-R1.2-01M-1*	Metadata contains elements which hold provenance information and can be mapped to PROV.
	FsF-R1.2-01M-2	Metadata contains provenance information using formal provenance ontologies (PROV-O).
FsF-R1.3-01M	*FsF-R1.3-01M-1*	Community specific metadata standard is detected using namespaces or schemas found in provided metadata or metadata services outputs.
	FsF-R1.3-01M-2	Community specific metadata standard is listed in the re3data record of the responsible repository.
	FsF-R1.3-01M-3	Multidisciplinary but community endorsed metadata (RDA Metadata Standards Catalog, fairsharing) standard is listed in the re3data record or detected by namespace.
FsF-R1.3-02D	*FsF-R1.3-02D-1*	The format of a data file given in the metadata is listed in the long term file formats, open file formats or scientific file formats controlled list.

	FsF-R1.3-02D-1a	The format of the data file is an open format.
	FsF-R1.3-02D-1b	The format of the data file is a long term format.
	FsF-R1.3-02D-1c	The format of the data file is a scientific format.

Table 2.2: Tests used by FUJI to evaluate the datasets with their appropriate descriptions.

There also are other tools and approaches when it comes to assessing the FAIRness of a given data record. One of them is FAIRshake [CWJ+19], for instance. It is a tool that has been created to allow for FAIRness evaluation based on different sets of metrics—it enables the communities of a certain field to create their own FAIRness assessment criteria. Another similar tool has been presented in [ABJ22]—O'FAIRe. It is a project based on the work of Garijo and Poveda-Villalón [GPV20, PVEAGC20] that merges several approaches of measuring the FAIRness of semantic resources. Yet another similar tool—FAIR-Checker [GRDL+23]—uses a certain set of SPARQL queries and SHACL constraints for the evaluation task.

There are also solutions that go beyond the concept of FAIRness applied to the metadata of records in a certain repository. There is a tool—Ontology Pitfall Scanner for FAIR principles (FOOPS!) [GCPV21] that can be used to establish the FAIRness of given vocabularies. It assesses the FAIRness of ontologies based on OWL or SKOS using a set of 24 metrics.

In this thesis, we chose to conduct our research using FUJI, as it assesses data records based on only certain RDA recommendations compared to O'FAIRe, FAIRshake or FAIR-Checker. Moreover, it is perfectly suited for our analyses—establishing FAIRness of certain research data records—unlike FOOPS! that evaluates vocabularies.

2.3 FAIRness of representative records

The commonly used research data repositories such as the Dryad Digital Repository[10], the Open Science Framework[11] or Zenodo[12] all store and export the metadata of their records. However, if one considered the automatic tool FUJI (mentioned in Section 2.2) as an oracle, the FAIRness level of that metadata varies from record to record and certain FAIR principles are often not satisfied. For example, the FAIRness score of the source code of FUJI published on Zenodo[13] is 66%—which is mainly the result of the complete lack of data description in the metadata. The metadata of this particular repository completely

[10] https://datadryad.org/
[11] https://osf.io/
[12] https://zenodo.org
[13] https://zenodo.org/records/4063720

violates the *FsF-A1-03D*[14], *FsF-R1.3-02D*[15], *FsF-R1-01MD-2*[16], *FsF-R1-01MD-3*[17] and *FsF-R1-01MD-4*[18] FAIRsFAIR metrics. Additionally, the *FsF-R1.2-01M-2*[19] metric is unsatisfied, which might suggest that the metadata regarding the history and record contributors is not modeled using one of the globally known ontologies or does not exist.

Even if the score is higher, certain key FAIRsFAIR metrics regarding data description in the metadata remain unsatisfied. An exemplary record[20] published on the Dryad Digital Repository achieves a score of 83%, however, the metrics regarding data description and data provenance *FsF-R1-01MD-2*, *FsF-R1-01MD-3*, *FsF-R1-01MD-4* and *FsF-R1.2-01M-2* are violated.

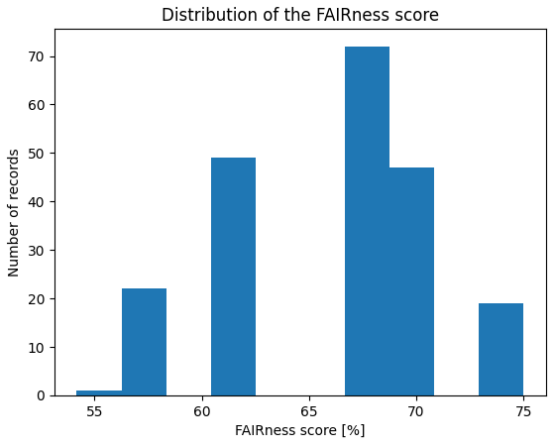

Figure 2.3: The distribution of FAIRness score (in %) of records in the TU Wien Research Data measured by the FUJI tool.

We have tested 210 records from the test instance of TU Wien Research Data using the automated tool FUJI to gain more concrete insight about their FAIRness. As illustrated in Figure 2.3, the score ranges from 54.17% (1 record) to 75% (19 records). The highest amount of records—72—achieved 66.67%, followed by 49 records with 62.50% and 47 records with 70.83%. One of the goals of this thesis is to find out how to improve these scores.

[14]Data is accessible through a standardized communication protocol.
[15]Data is available in a file format recommended by the target research community.
[16]Verifiable data descriptors.
[17]Data content matches file type and size or protocol specified in metadata.
[18]Data content matches measured variables or observation types specified in metadata.
[19]Metadata contains provenance information using formal provenance ontologies.
[20]`https://datadryad.org/stash/dataset/doi:10.5061/dryad.k3j9kd5hz`

2.4 FAIR Digital Objects

The GO FAIR[21] initiative aims to coordinate existing FAIR implementations. One of the frameworks recognized by GO FAIR are FAIR Digital Objects specified by the FDO Forum[22] [ICD+23]. This relatively novel concept is based on the already existing notion of Digital Objects introduced by Kahn and Wilensky in 2006 [KW06].

A FAIR Digital Object is defined as a *unit composed of data that is a sequence of bits, [...] interpretable by one or more computer systems, and having as essential elements an assigned globally unique and persistent identifier (PID), a type definition for the object as a whole and a metadata description [...], making the whole findable, accessible, interoperable and reusable both by humans and computers* [CML+25]. The FDO specification is completely built around the FAIR principles and it allows a seamless integration with a global LOD vocabulary (such as schema.org). Moreover, the specification strictly requires a precise definition of not only the Digital Object itself but also of all its attributes and their types. Additionally, an FDO has to be accessible through a clearly defined system which is capable of resolving all the PIDs that are included in the object's metadata and attributes, which satisfies the conditions of LOD.

The FDO specification defines a FAIR Digital Object as an entity that has a number of descriptive properties. Those are its attributes. These attributes can describe anything related to the object, however there are two specific kinds of attributes: *type* and *profile*. The first serves a purpose of a description about the entity and can be just a human-readable string. The latter one is slightly more important—it provides a machine-actionable representation of how a certain FDO can be automatically parsed. Additionally, all attributes have to have a machine-readable description—an attribute definition—that also is an FDO. How the elements of the FDO specification are connected to each other on the conceptual level is rather complicated and could be summed up in a single principle: *everything is an FDO*. We explain this in further detail in Section 3.2. The relationships between the entities of an FDO system have been depicted in Figure 2.4.

[21] https://www.go-fair.org/
[22] https://fairdo.org/

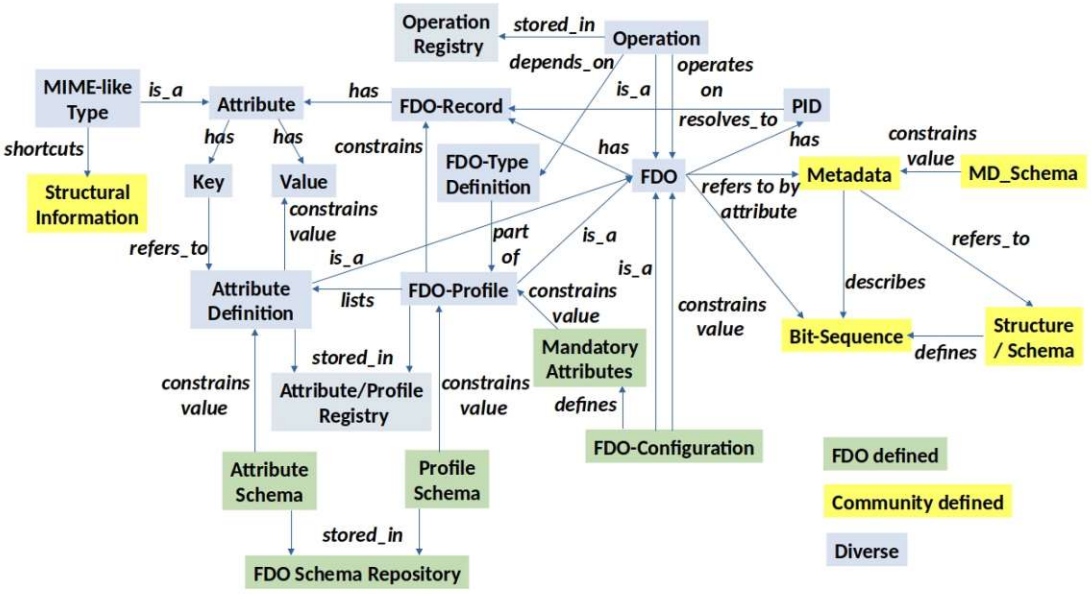

Figure 2.4: A diagram of the elements of an FDO system from the FDO specification [CML+25].

2.5 Cordra

The Cordra open source Digital Object management software is a tool developed and maintained by the Corporation for National Research Initiatives (CNRI)[23]. It leverages elements of the Digital Object Architecture (DOA) specified by the DONA foundation[24]. It is packed with customization possibilities, such as the option to define custom object validation mechanisms [Cor23]. Furthermore, the objects and their identifiers saved in an instance of the Cordra software are accessible through well-specified protocols—DOIP and DOIRP [KBL+18, KBTS+22]—which already satisfies the *FsF-A1-02M* accessibility FAIRsFAIR metric. These qualities make it a suitable backbone for a potential FDO framework [TS22]. Additionally, the objects managed by Cordra are JSON records and payloads, which allows for a straightforward incorporation of JSON-LD LOD annotations into the said framework.

Other digital object management software such as the Fedora Repository[25] or DSpace[26] exist, however these are heavyweight solutions more suitable for more involved and complex tasks [Fay10]. Cordra offers a lightweight and highly customizable digital object storage with an efficient backend and a user-friendly frontend, which has been depicted in Figure 2.5.

[23] https://www.cnri.reston.va.us/
[24] https://www.dona.net/digitalobjectarchitecture
[25] https://fedorarepository.org/
[26] https://dspace.org/

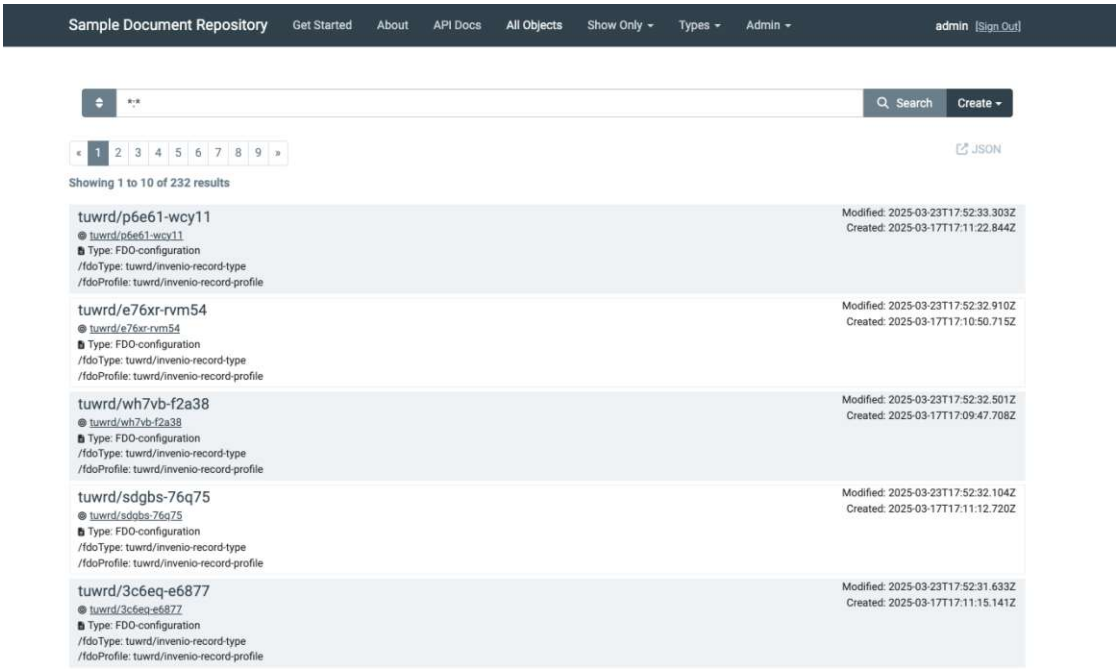

Figure 2.5: A screenshot of the Cordra user interface.

2.6 InvenioRDM

Research data repositories are platforms where researchers can deposit data and also create appropriate metadata that describes this data. It is crucial that the repositories guarantee a dependable level of accessibility of their records and facilitate a straightforward data sharing process using Digital Object Identifiers (DOIs). Nowadays, a reliable data repository is a key aspect in the Research Data Management Services stack [QT25].

One of the many notable examples of such a repository is the PANGAEA[27] information system [FMS+23]. It is an open access repository for georeferenced data from environmental sciences. It focuses on FAIR and open data infrastructures to facilitate for an easy data exchange among researchers. Another example is the Dryad Digital Repository[28], which is not only an open data publishing platform but is also developed by a community that strives to enable the re-use of all research data.

The latter example was originally based on the DSpace repository software, developed by the Massachusetts Institute of Technology and Hewlett-Packard. DSpace is an open source framework that can be used for many purposes, one of which is creating open access repositories for scholarly content. InvenioRDM is another such framework. It is an open source collaboration between many scientific institutions, including CERN,

[27] https://www.pangaea.de/
[28] https://datadryad.org/

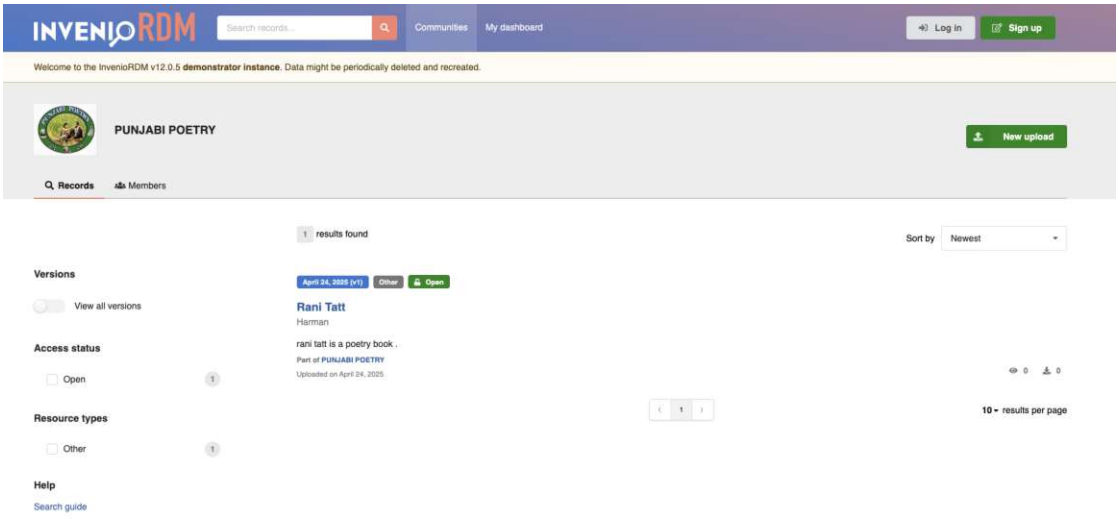

Figure 2.6: A screenshot of the InvenioRDM user interface.

the Northwestern University and many more[29]. It is a representative example of how a research data repository framework should be structured, as it has been adopted by numerous research data repositories, such as Zenodo and TU Wien Research Data. Moreover, the community behind this framework aims to make the repository content more machine-accessible. Following recent developments, InvenioRDM is supposed to support the FAIR principles [Car23] and Signposting [Vig24]. It supports assignment of Digital Object Identifiers via DataCite, multiple metadata formats and provides both a web interface (depicted in Figure 2.6) and APIs (REST, OAI-PMH) for discovery of its records.

2.7 Discussion

In this chapter, we discussed in more detail the concepts and technologies that are closely related to our work. We provided an in-depth background about the FAIR principles themselves and focussed on the FAIR recommendations and projects that are relevant to this thesis. We have expanded upon the concept of FAIRness and introduced more information about the FAIRness indicators, especially those based directly on the RDA recommendations—the FAIRsFAIR metrics. We have presented these 17 metrics and explained how they relate to the FAIR principles. We also discussed the existing FAIRness assessment tools—FAIRshake, O'FAIRe, FAIR-Checker, FOOPS! and FUJI. For the purpose of this thesis, we chose to use the latter one, since it assesses data records following an approach based solely on certain RDA recommendations unlike the other tools.

[29] https://invenio-software.org/people/

Moreover, we also discussed the FAIR Digital Objects and provided more insight about this particular FAIR recommendation. Furthermore, we presented an overview of the technologies that we analyzed in search for tools suitable for implementation of our data repository system based on FDOs. We chose to base our solution on the InvenioRDM project, as it is a representative example of how a research data repository framework should be structured—it has been adopted by numerous research data repositories, such as Zenodo and TU Wien Research Data. We also selected the Cordra open source Digital Object management software as a backend for hosting the FDOs due to it being lightweight and highly customizable.

CHAPTER 3

Proposed solution

In this thesis, we aim to propose a conceptual architecture of an extended research data repository that represents the metadata of its records as FDOs. This extended research data repository combines two services—a research data repository that holds its records' data and an FDO backend service, where the metadata of the records will be stored as FDOs.

To store the metadata in the FDO backend, we need to first choose an appropriate service that could be used to store FDOs. Secondly, we need to design an appropriate FDO infrastructure that allows us to model records' metadata as FDOs.

For this purpose, we have selected the Cordra open source Digital Object management software as our FDO backend. It is a suitable service for such a backend due to it being lightweight and highly customizable. Furthermore, it is based on the Digital Object Architecture (DOA), so the objects it holds are accessible through well-specified protocols.

Moreover, we designed a concrete FDO infrastructure that has been created by strictly following the FDO specification [CML+25]. We have split this effort into two parts. Firstly, we designed an FDO framework that consists of 6 base building blocks—we called them core FDOs. This framework can be fine-tuned and reused for purposes different from the one that we consider in this thesis—representation of the metadata of records from a research data repository. Secondly, we created an extension on top of this framework that allows us to represent records' metadata. We called this extension the *Research Metadata Extension*. It is a set of additional FDOs that adjust the aforementioned framework to our needs. We present in Figure 3.1 an overview how these elements relate to each other.

As another key aspect of this thesis, we prepared a procedure that we called migration. It is a conceptualization of how the records from an already existing research data repository should be migrated to an FDO backend without damaging or losing any data. Furthermore, it is straightforward to adopt and implement alongside an already existing system.

Our solution is intended to be used by an administrator of an already existing research data repository, who intends to elevate the FAIRness of records in that repository. This person can use the migration procedure that we propose to migrate the metadata of records from their repository to our novel FDO framework. It is fairly straightforward and should not require a significant amount of manual labor. We provide more implementation details and information about how our solution influences the said FAIRness in Chapter 4.

In this chapter, we present the rationale behind our choice of the FDO backend along with the conceptual architecture of the extended research data repository we propose—we expand upon this in Section 3.1. In Section 3.2, we present our novel FDO framework and, in Section 3.3, we present our extension of this framework. At the end of this chapter, in Section 3.4, we also present the migration procedure.

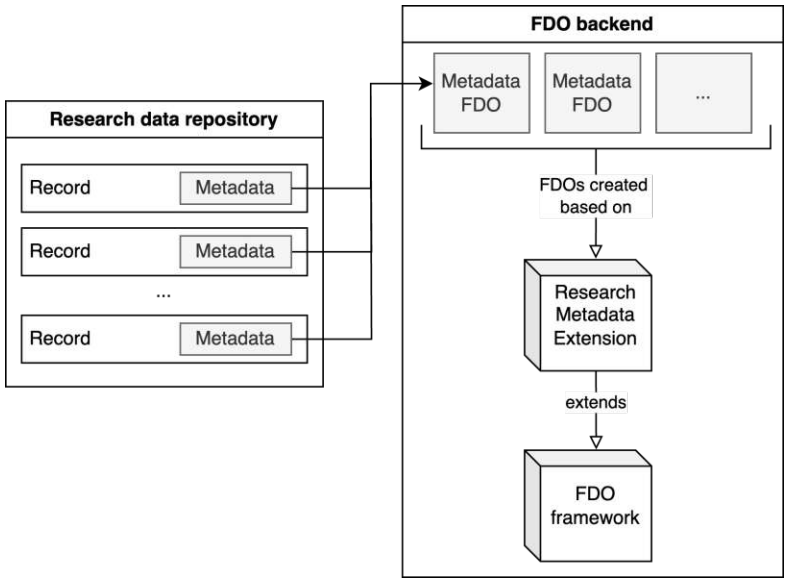

Figure 3.1: A diagram with a conceptualization of the proposed solution. On the left side is a standard research data repository that contains both data records and their metadata. On the right side, we present how the metadata is represented in a FDO backend based on our FDO framework and the *Research Metadata Extension*.

3.1 Solution architecture

We aim to propose a conceptual architecture of an extended research data repository that represents the metadata of its records as FDOs. This extended research data repository combines two services—a research data repository that holds its records' data and an FDO backend service, where the metadata of the records will be stored as FDOs. We present this in Figure 3.2, where we juxtapose a standard research data repository with the extended one that we propose.

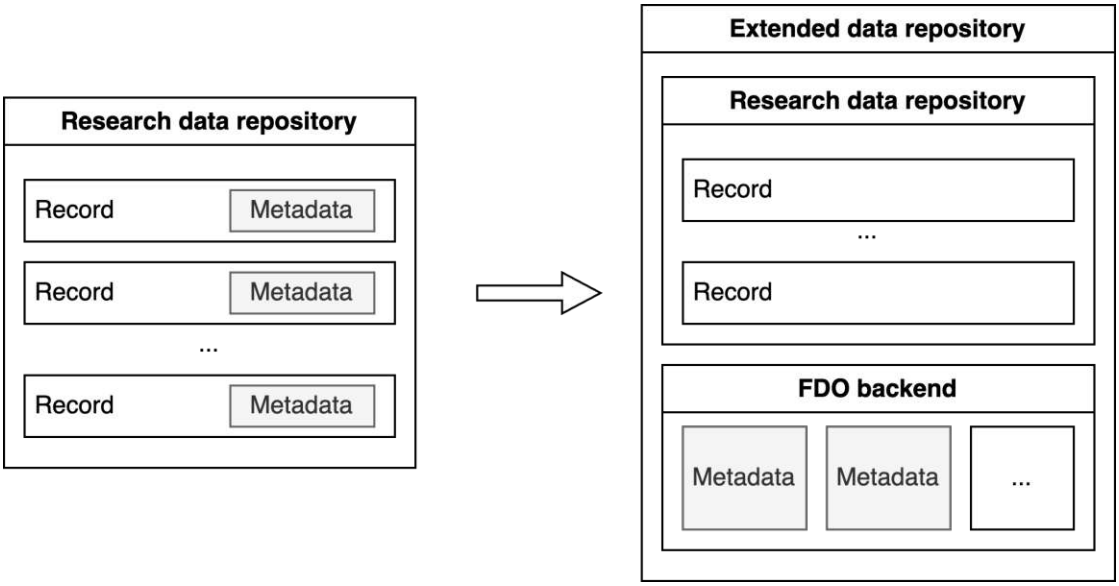

Figure 3.2: A diagram with a conceptualization of the proposed solution. On the left side is a standard research data repository that contains both data records and their metadata. On the right side, we present an extended data repository that stores the metadata of its records in an FDO backend as FDOs.

For the realization of this architecture, we need to choose an appropriate service for the FDO backend and implement a translation layer between the two services—the research data repository and the FDO backend. We present an example of the latter one in Chapter 4 and we provide a rationale behind the chosen FDO backend below.

An ideal service for an FDO backend would be a lightweight application. We are not looking for a multipurpose tool which can be used as a digital library but a service that has a single purpose—storing digital entities. This service would also need to have at least some amount of possible customization when it comes to the structure of the objects it stores. It would also be an advantage, if the service exported its objects in a well-known format, such as JSON or XML.

The Cordra open source Digital Object management software seems to be a suitable candidate for our needs. It is a tool developed and maintained by the Corporation for National Research Initiatives (CNRI)[1]. It is a lightweight application for storing digital entities and is packed with customization possibilities, such as the option to define custom object validation mechanisms [Cor23]—exactly what we require. As an additional advantage, it leverages elements of the Digital Object Architecture (DOA) specified by the DONA foundation[2]. That means that the objects and their identifiers saved in an instance of the Cordra software are accessible through well-specified protocols—DOIP

[1] https://www.cnri.reston.va.us/
[2] https://www.dona.net/digitalobjectarchitecture

and DOIRP [KBL+18, KBTS+22]—which already satisfies the *FsF-A1-02M* accessibility FAIRsFAIR metric. These qualities make it a suitable backbone for a potential FDO framework [TS22]. Additionally, the objects managed by Cordra are JSON records and payloads, which allows for a straightforward incorporation of JSON-LD LOD annotations into the said framework.

Other digital object management software such as the Fedora Repository[3] or DSpace[4] exist, however these are heavyweight solutions more suitable for more involved and complex tasks [Fay10]. Cordra offers a lightweight and highly customizable digital object storage with all the necessary functionalities for an FDO backend.

3.2 FDO framework

The FDO specification defines a FAIR Digital Object as a *unit composed of data that is a sequence of bits, or a set of sequences of bits* [CML+25] which is accompanied by an FDO record—a set of key-value pairs called attributes. In the FDO framework constructed for the purpose of this thesis, the FDO record and the FDO itself are a singular entity. This is *the most straightforward and thus the default way* [CML+25] to realize an FDO system, which is also possible in the chosen Digital Objects backend—Cordra.

The main goal of an FDO system is to make all of its components fully interpretable by machines. This means that not only the meaning of a singular entity should be made understandable to computers but also the entirety of its components, which, in case of the FDO record, means its attributes. Therefore, each attribute has to have a description—a so-called *attribute definition*—that enables its machine-actionability. This description, to make the whole system bulletproof, also must be machine-actionable and has to be an FDO as well. The definition of such a system can be, thus, summed up in a single principle: *everything is an FDO*.

Each FDO has two important kinds of attributes: type and profile. The first one is supposed to be a description of how computers are able to process the contents of a certain entity. In this particular framework, a type is an FDO that contains a human-readable string, which is meant to serve the purpose of the intended description. The profile is a slightly more involved attribute. Its function is to explicitly model the structure of the FDO it describes. From a perspective of an automatic FDO parser, the profile is an entry point to gain necessary information about what attributes are to be expected while processing an FDO that this profile describes. In this FDO framework, a profile is used to resolve attribute keys of a certain FDO and validate their values against the appropriate attribute definitions, which is explained in more detail further in this section.

In order to distinguish the aforementioned important kinds of attributes in the proposed FDO framework, the FDO structure should be standardized and a set of required attributes for each FDO should be distinguished. Following the FDO specification, only

[3] https://fedorarepository.org/
[4] https://dspace.org/

the profile has to be a mandatory attribute, however an implementation is allowed to require other attributes in an FDO as well. The mandatory attributes of this FDO framework have been defined in a Cordra Schema object, which intuitively can be viewed as a Cordra-type of FDOs. It has been called *FDO-configuration*[5] and it requires three attributes:

- *id*—a PID of a certain entity, shared between the targeted research data repository and Cordra[6];
- *fdoType*—a reference to an FDO which describes the type of this FDO[7];
- *fdoProfile*—a reference to an FDO that is this FDO's profile[7].

The FDO-configuration definition is similar to a JSON schema with an addition of Cordra elements. This means that newly created objects of the FDO-configuration type undergo a validation procedure that resembles JSON schema validation. However, thanks to the Cordra customization possibilities, this step can be extended. For the purpose of this framework, the *beforeSchemaValidation* Cordra hook has been overwritten with an additional validation mechanism, which is described in more detail further in this section.

The FDO framework proposed in this thesis consists of 6 core FDOs:

1. FDO-content-profile,
2. FDO-type-profile,
3. FDO-attribute-definition-type,
4. FDO-profile-type,
5. FDO-schema-attribute-definition and
6. FDO-description-attribute-definition.

Given the fact that all of the above are the same objects at the conceptual level—all are FDOs and follow the FDO-configuration schema—but are created with different purposes—some of them are attribute definitions, some are profiles and some are types—it might become difficult to tell them apart. For this reason, all objects in this framework follow the following naming convention:

[5]This naming convention is in a way compliant with the recommendations of the FDO specification, where FDO-configuration is an object that introduces mandatory attributes of a profile. This FDO framework simply assumes that all FDOs (not only profiles) share a common base structure.

[6]A convenience attribute to easily identify an object in the Cordra system.

[7]Because circular dependencies exist in the system, these fields can also be blank JSON objects. This is necessary for certain objects to be created first, so that the references could be added afterwards.

- the names of the core FDOs begin with the *FDO-* prefix,
- the names of types end with the *-type* suffix,
- the names of profiles end with the *-profile* suffix and
- the names of attribute definitions end with the *-attribute-definition* suffix.

Figure 3.3 depicts how the core FDOs are structured and connected to each other. The FDO-content-profile is the primary profile of attribute definitions and profiles, including itself. Apart from the required attributes: *id*, *fdoType* and *fdoProfile*, it has a *schema* attribute, which is a JSON schema. This *schema* attribute contains a definition of the object that the profile is describing and is used to validate this object before its creation. It might contain references to other FDOs (for example to attribute definitions), which have to be resolved prior to the validation itself. Therefore, the object's validation procedure is structured as follows:

1. resolve object's profile reference from the mandatory *fdoProfile* attribute[8];
2. if the object's profile has an attribute called *schema*, resolve possible references to other FDOs in the value of that attribute and perform schema validation of the object against the resolved schema;
3. if the object itself has a *schema* attribute and it contains references to other FDOs, check if the referenced objects exist. If not, fail.

During the resolution step (2), a parser looks for a specifically structured string that contains a reference to other FDOs in the *schema* attribute of a given profile. This string is then replaced with the value of the *schema* attribute found in the referenced object, which, similarly to the *schema* attribute of the profile itself, is also a JSON schema. Afterwards, the value of the resolved *schema* attribute in the profile is a JSON schema that can be used to validate a certain object. An example of the schema resolution is presented in the listings: 3.1, 3.2, 3.3. The reference string *FDO_REF:CORDRA_PREFIX/FDO-schema-attribute-definition* is being replaced with *{"type": "object"}* during the resolution step, which is visible in Listing 3.3.

[8]During creation of the 6 core FDOs, the validation is turned off—these objects are first created without the profile and type references, which are added later, when all primary objects exist in the system.

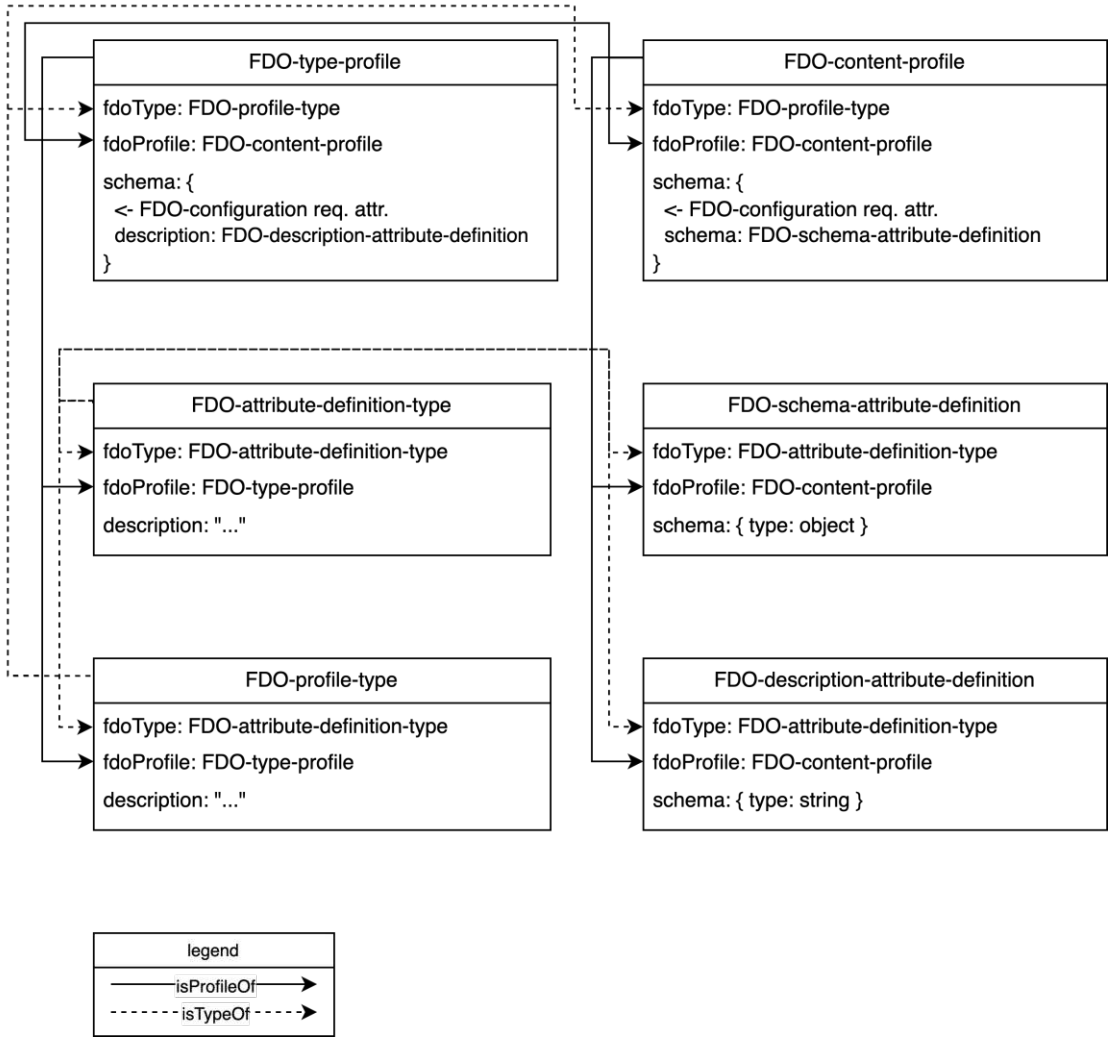

Figure 3.3: A diagram of the proposed FDO framework depicting the core FDOs with their attributes. For the purposes of brevity, the convenience attribute *id* has been left out and an abbreviation: *req. attr.* has been used to annotate: *required attributes.*

In case of the FDO-content-profile, the *schema* attribute describes an object that contains the mandatory fields inherited from the FDO-configuration and a *schema* attribute, which is defined by the FDO-schema-attribute-definition. This attribute definition has the *schema* attribute as well, which, is a JSON schema that describes a singular entity of JSON-type *object*. Thus, any object that is described by the FDO-content-profile must have the required FDO-configuration attributes and a *schema* attribute, the value of which is a JSON object. This is true for the FDO-content-profile itself and therefore it is its own profile.

FDO-content-profile is also the profile of FDO-type-profile, FDO-schema-attribute-

Listing 3.1: The FDO-content-profile object.

```
{
  "id": "FDO-content-profile",
  "fdoType": "CORDRA_PREFIX/FDO-profile-type",
  "fdoProfile": "CORDRA_PREFIX/FDO-content-profile",
  "schema": {
    "allOf": [
      {"$ref": "FDO-configuration"},
      {
        "type": "object",
        "properties": {
          "schema": "FDO_REF:CORDRA_PREFIX/FDO-schema-attribute-
              definition"
        },
        "required": ["schema"]
      }
    ],
    "unevaluatedProperties": false
  }
}
```

Listing 3.2: The FDO-schema-attribute-definition object.

```
{
  "id": "FDO-schema-attribute-definition",
  "fdoType": "CORDRA_PREFIX/FDO-attribute-definition-type",
  "fdoProfile": "CORDRA_PREFIX/FDO-content-profile",
  "schema": {"type": "object"}
}
```

Listing 3.3: The schema attribute of the FDO-content-profile after resolution.

```
"schema": {
  "allOf": [
    {"$ref": "FDO-configuration"},
    {
      "type": "object",
      "properties": {
        "schema": {"type": "object"}
      },
      "required": ["schema"]
    }
  ],
  "unevaluatedProperties": false
}
```

definition and FDO-description-attribute-definition. All of these FDOs have exactly the attributes specified by the FDO-content-profile and use the *schema* attribute to provide a machine-interpretable description of other elements of this system. The FDO-type-profile FDO is supposed to be a profile that provides a description of type-FDOs. Its *schema* attribute allows for an attribute called *description*, which is specified by the FDO-description-attribute-definition as a *string*. The description attribute is supposed to be a human-readable description of the objects it describes. FDO-type-profile is the profile of FDO-profile-type and FDO-attribute-definition-type, which are the types of other profiles and attribute-definitions respectively. Introducing these type-FDOs might seem superfluous, the *fdoType* attribute could just have been a simple string, which would also be enough to provide a human-readable description of a certain entity. However, these objects have been created to elevate the machine-interpretability of the system as a whole by making it possible for machines to understand the value of the *fdoType* attribute based on another FDO that describes it.

Together, these 6 core FDOs are the primary building blocks of the FDO framework presented in this thesis. This construction of profiles and attribute definitions ensures machine-interpretability of the entire system. The objects are defined by the *schema* attributes of their profiles and objects' attributes are defined by the *schema* attributes of their attribute definitions. Given the fact that the *schema* attribute is a machine-readable JSON schema, the whole system is *understandable* to a potential automated parser.

We present these core FDOs along with the FDO-configuration Cordra Schema in Section 3.2.1.

3.2.1 Core FDOs

We present the core FAIR Digital Objects and the FDO-configuration Cordra Schema in the listings provided in this section along with an appropriate description.

FDO-configuration

The Listing 3.4 presents the FDO-configuration Cordra Schema and the custom validation procedure is depicted in Listing 3.5. This Cordra-structure is a schema that every FDO from the framework must follow.

Listing 3.4: The FDO-configuration Cordra schema.

```
{
    "type": "object",
    "required": [
        "fdoType",
        "fdoProfile",
        "id"
    ],
    "definitions": {
        "attribute-metatype": {
        "oneOf": [
            {"type": "object"},
            {
                "type": "string",
                "cordra": {
                    "type": {
                    "handleReference": {
                        "types": ["FDO-configuration"]
                    }
                    },
                    "preview": {"showInPreview": true}
                }
            }
        ]
        }
    },
    "properties": {
        "id": {
            "type": "string"
        },
        "fdoType": {
            "title": "FDO type",
            "$ref": "#/definitions/attribute-metatype"
        },
        "fdoProfile": {
            "title": "FDO Profile",
            "$ref": "#/definitions/attribute-metatype"
        }
    }
}
```

Listing 3.5: The FDO-configuration validation procedure.

```
const cordraUtil = require('cordra-util');
const cordra = require('cordra');

exports.methods = {};
exports.methods.validate = beforeSchemaValidation;
exports.beforeSchemaValidation = beforeSchemaValidation;

const FDO_REF_PREFIX = "FDO_REF"

function resolveFdoReferences(obj) {
    if (typeof obj !== 'object' || obj === null) return;

    // try to resolve all the strings
    for (let prop in obj) {
        if (typeof obj[prop] !== 'string') { continue; }
        if (!obj[prop].includes(`${FDO_REF_PREFIX}:`)) {continue; }

        let referencedObjectId = obj[prop].replace(`${FDO_REF_PREFIX}:`,
            '');
        let referencedObject = cordra.get(referencedObjectId);

        if (referencedObject == null) { throw `could not resolve
            reference: ${obj[prop]}` }
        obj[prop] = referencedObject.content.schema;
    }

    for (let prop in obj) { resolveFdoReferences(obj[prop]) }
}

function beforeSchemaValidation(object, context) {
    // NOP for special objects
    if (typeof object.content.fdoProfile !== 'string') {
        return object;
    }

    // retrieve FDO's profile
    let profile = cordra.get(object.content.fdoProfile);
    if (profile == null) {
        throw `unknown profile: ${object.content.fdoProfile}`;
    }

    let schema = JSON.parse(JSON.stringify(profile.content.schema));
    resolveFdoReferences(schema);

    let rsp = cordraUtil.validateWithSchema(object.content, schema);
    if (rsp.success == false) {
        throw `object definition is not valid with regard to its profile:
            ${rsp.errors[0].message}`;
    }

    // if the object contains a schema, validate the schema
    if (object.content.schema) {
```

```
            schema = JSON.parse(JSON.stringify(object.content.schema));
            resolveFdoReferences(schema);
        }

        return object;
    }
```

FDO-content-profile

Listing 3.6 represents our implementation of the FDO-content-profile. It is defined as an instance of the FDO-configuration Cordra Schema, which makes it a FAIR Digital Object. This is the most important profile in the framework. It is used as the value of the *fdoProfile* attribute of most core FDOs, such as: FDO-content-profile (itself), FDO-type-profile, FDO-schema-attribute-definition and FDO-description-attribute-definition.

Listing 3.6: The FDO-content-profile FAIR Digital Object.

```
{
    "id": "FDO-content-profile",
    "fdoType": "CORDRA_PREFIX/FDO-profile-type",
    "fdoProfile": "CORDRA_PREFIX/FDO-content-profile",
    "schema": {
      "allOf": [
        {
          "$ref": "FDO-configuration"
        },
        {
          "type": "object",
          "properties": {
            "schema": "FDO_REF:CORDRA_PREFIX/FDO-schema-attribute-
                definition"
          },
          "required": [
            "schema"
          ]
        }
      ],
      "unevaluatedProperties": false
    }
  }
```

FDO-type-profile

Listing 3.7 represents our implementation of the FDO-type-profile. It is defined as an instance of the FDO-configuration Cordra Schema, which makes it a FAIR Digital Object. This is the profile-FDO of type-FDOs in our framework. It is the value of the *fdoType* attribute for FDO-attribute-definition-type and FDO-profile-type.

Listing 3.7: The FDO-type-profile FAIR Digital Object.

```
1   {
2       "id": "FDO-type-profile",
3       "fdoType": "CORDRA_PREFIX/FDO-profile-type",
4       "fdoProfile": "CORDRA_PREFIX/FDO-content-profile",
5       "schema": {
6         "allOf": [
7           {
8             "$ref": "FDO-configuration"
9           },
10          {
11            "type": "object",
12            "properties": {
13              "description": "FDO_REF:CORDRA_PREFIX/FDO-description-
                    attribute-definition"
14            },
15            "required": [
16              "description"
17            ]
18          }
19        ],
20        "unevaluatedProperties": false
21      }
22    }
```

FDO-attribute-definition-type

Listing 3.8 represents our implementation of the FDO-attribute-definition-type. It is defined as an instance of the FDO-configuration Cordra Schema, which makes it a FAIR Digital Object. This is the type-FDO for attribute-definition-FDOs, such as: FDO-profile-type, FDO-schema-attribute-definition and FDO-description-attribute-definition.

Listing 3.8: The FDO-attribute-definition-type FAIR Digital Object.

```
1    {
2        "id": "FDO-attribute-definition-type",
3        "fdoType": "CORDRA_PREFIX/FDO-attribute-definition-type",
4        "fdoProfile": "CORDRA_PREFIX/FDO-type-profile",
5        "description": "This is an attribute definition used in an FDO
             framework that follows the FDO specification"
6      }
```

FDO-profile-type

Listing 3.9 represents our implementation of the FDO-profile-type. It is defined as an instance of the FDO-configuration Cordra Schema, which makes it a FAIR Digital Object. This is the type-FDO for profile-FDOs. It is the value of the *fdoType* attribute for: FDO-content-profile and FDO-type-profile.

Listing 3.9: The FDO-profile-type FAIR Digital Object.

```
1    {
2        "id": "FDO-profile-type",
3        "fdoType": "CORDRA_PREFIX/FDO-attribute-definition-type",
4        "fdoProfile": "CORDRA_PREFIX/FDO-type-profile",
5        "description": "This is a profile used in an FDO framework that
             follows the FDO specification"
6    }
```

FDO-schema-attribute-definition

Listing 3.10 represents our implementation of the FDO-schema-attribute-definition. It is defined as an instance of the FDO-configuration Cordra Schema, which makes it a FAIR Digital Object. This is the definition of the *schema* attribute, which appears in most descriptive FDOs, such as profiles and attribute definitions. It is defined as a JSON object, which makes the descriptive elements quite flexible—their *schema* attribute can contain any valid JSON object.

Listing 3.10: The FDO-schema-attribute-definition FAIR Digital Object.

```
{
    "id": "FDO-schema-attribute-definition",
    "fdoType": "CORDRA_PREFIX/FDO-attribute-definition-type",
    "fdoProfile": "CORDRA_PREFIX/FDO-content-profile",
    "schema": {
      "type": "object"
    }
}
```

FDO-description-attribute-definition

Listing 3.11 represents our implementation of the FDO-description-attribute-definition. It is defined as an instance of the FDO-configuration Cordra Schema, which makes it a FAIR Digital Object. This is the definition of the *description* attribute, which appears in every type-FDO that follows the FDO-type-profile. It is defined as a single string, which allows us to provide a human-readable definition of a certain FDO.

Listing 3.11: The FDO-description-attribute-definition FAIR Digital Object.

```
1    {
2        "id": "FDO-description-attribute-definition",
3        "fdoType": "CORDRA_PREFIX/FDO-attribute-definition-type",
4        "fdoProfile": "CORDRA_PREFIX/FDO-content-profile",
5        "schema": {
6          "type": "string"
7        }
8    }
```

3.3 Research Metadata Extension

To be able to model the metadata of records from a research data repository in our FDO backend, the FDO framework from Section 3.2 has to be extended by a set of profiles, attribute-definitions and types. We present this set of FDOs in this section and call it the *Research Metadata Extension*.

These FDOs have been designed in a way that allows a straightforward integration with Linked Open Data (LOD) vocabularies of choice, which elevates the FAIRness of the FDOs that are going to be stored in our FDO backend. We present, in Section 3.3.1, how this integration is achieved using our extension and, in Section 3.3.2, which vocabularies we chose to use.

Furthermore, we designed a set of attributes that the FDOs representing research data records' metadata should have. We present this set of attributes in Section 3.3.3. Finally, as a product of the work and decisions described in Section 3.3.1, Section 3.3.2 and Section 3.3.3, we present the set of FDOs that the *Research Metadata Extension* consists of. In Section 3.3.4 each of these FDOs is presented with an appropriate description.

3.3.1 Integration with Linked Open Data

The FDOs themselves do not incorporate any particular LOD structures. Any LOD annotations are supposed to be included independently of the core FDOs described in Section 3.2. For this purpose a certain set of attributes should be used. This way, the framework itself can be reused for other purposes.

The objects hosted in the Cordra software can all be exported as JSON objects and thus the LOD integration can be carried out using attributes that follow JSON-LD naming conventions. This thesis proposes to use attributes named *@context*, *@type* and *@id* to introduce LOD annotations and vocabularies.

An example profile-FDO that allows for integration with LOD vocabularies is presented in Listing 3.12. It is actually a fragment of the profile that has been used to represent the metadata of records from a research data repository as FDOs in the proposed FDO framework. It introduces the *@context* and *@type* attributes that are described by object-attribute-definition and text-attribute-definition FDOs. These attribute definitions refer to JSON structures of type object and string and have been introduced in Section 3.3.4.

An example FDO that is described by this fragment of invenio-record-profile is presented in Listing 3.13. It uses 3 different vocabularies in its *@context*: schema.org, PROV-O and DCMI Metadata Terms. This FDO can be easily parsed to a set of RDF triples with meaningful predicates, which makes it understandable in the context of LOD. It is also important to note that the names of the attributes do not have to conform with the vocabularies, as their meaning gets extended in the *@context* attribute. For example, the *identifier* attribute is not mapped to the schema.org vocabulary on its own. This is done in the *@context* attribute by introducing the mapping *"identifier": "schema:identifier"*. The name of the attribute (and its counterpart in the mapping) could

Listing 3.12: The abridged snippet of the invenio-record-profile FDO.

```
{
  "id": "invenio-record-profile",
  "fdoType": "CORDRA_PREFIX/FDO-profile-type",
  "fdoProfile": "CORDRA_PREFIX/FDO-content-profile",
  "schema": {
    "allOf": [
      {"$ref": "FDO-configuration"},
      {
        "type": "object",
        "properties": {
          "@context": "FDO_REF:CORDRA_PREFIX/object-attribute-definition",
          "@type": "FDO_REF:CORDRA_PREFIX/text-attribute-definition",
          "identifier": "FDO_REF:CORDRA_PREFIX/text-attribute-definition",
          "wasDerivedFrom": "FDO_REF:CORDRA_PREFIX/text-attribute-definition
              ",
          "isVersionOf": "FDO_REF:CORDRA_PREFIX/text-attribute-definition"
        }
      }
    ]
  }
}
```

be changed to something different, say *id*, and it would still be resolved as a reference to *"schema:identifier"*. This illustrates, that the attribute names are independent of the chosen LOD vocabularies and in case of any conceptual changes to the chosen vocabulary in a single FDO, its profile does not necessarily need to be altered—it suffices to change the value of its *@context* attribute.

3.3.2 Choice of the Linked Open Data vocabularies

Even though the framework is generally independent of LOD annotations, certain choices had to be made to be able to model the metadata of the records of a given research data repository. The goal here is to be able to export the FDOs as JSON-LD documents that are highly descriptive on their own. It makes the objects also understandable to tools that are not accustomed to processing FDOs, such as FUJI. Moreover, given the fact that FUJI is the chosen FAIRness assessment tool, the usage of proper vocabularies is the primary factor that influences the score.

The available LOD taxonomies vary in size and popularity. For example, the W3C recommends the usage of Simple Knowledge Organization System (SKOS) [MB09] and Data Catalog Vocabulary (DCAT) [CBB+24]. However, these are relatively small vocabularies with limited number of terms. There also exists a much broader vocabulary created in 2011 by Google, Yahoo, Yandex and Bing—schema.org[9], which, lately, has

[9] https://schema.org/

Listing 3.13: An example FDO that introduces 3 LOD vocabularies: schema.org, PROV-O and DCMI Metadata Terms in its @context attribute.

```
{
  "id": "p6e61-wcy11",
  "fdoType": "CORDRA_PREFIX/invenio-record-type",
  "fdoProfile": "CORDRA_PREFIX/invenio-record-profile",
  "@context": {
    "schema": "http://schema.org/",
    "prov": "http://www.w3.org/ns/prov#",
    "dcterms": "http://purl.org/dc/terms/",
    "identifier": "schema:identifier",
    "wasDerivedFrom": "prov:wasDerivedFrom",
    "isVersionOf": "dcterms:isVersionOf"
  },
  "@type": "https://schema.org/Dataset",
  "identifier": "https://doi.org/10.70124/p6e61-wcy11",
  "wasDerivedFrom": "https://test.researchdata.tuwien.at/records/p6e61-wcy11
      ",
  "isVersionOf": "10.70124/skx29-p7x96"
}
```

become one of the most popular vocabularies worldwide [GBM16, IASK25]. Nowadays, a large fraction of the structured data is based on schema.org annotations [TTHS19].

Numerous research data repositories export their records' metadata using a few different vocabularies. The set of these vocabularies varies and can be oftentimes customized, however this thesis focuses on the ones that are available in the TU Wien Research Data repository. This includes for example: schema.org, the DataCite Ontology[10] and DCMI Metadata Terms[11].

On the other side, the possibilities of the available FAIRness assessment tools also have to be taken into account. We expand upon this topic in Section 4.2 and Section 4.3. After a thorough review which vocabularies are considered FAIR both by the communities behind research data repositories and FAIRness assessment tools, we decided to use annotations from the following three vocabularies: schema.org, PROV-O[12] and DCMI Metadata Terms.

3.3.3 Metadata structure

Metadata of a record from a research data repository is intended to be modelled using the profile *invenio-record-profile*. We named it this way, given the fact that in our research, we focus on this particular repository framework (InvenioRDM), however the name itself can be changed. It allows for 22 attributes, two of which have to follow the JSON-LD

[10] https://lov.linkeddata.es/dataset/lov/vocabs/dcite
[11] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
[12] https://www.w3.org/TR/prov-o/

naming conventions: *@context* and *@type*. The rest can be named arbitrarily, however we chose names that resemble annotations from the intended vocabularies. Overall, there are 17 attributes that represent properties from schema.org, those are:

- *identifier*,
- *name*,
- *author*,
- *editor*,
- *publisher*,
- *keywords*,
- *datePublished*,
- *dateModified*,
- *inLanguage*,
- *contentSize*,
- *version*,
- *license*,
- *description*,
- *citation*,
- *url*,
- *distribution* and
- *conditionsOfAccess*.

Additionally, the FDOs can also be equipped with two attributes originating from DCMI Metadata Terms: *hasPart* and *isVersionOf* and one attribute—*wasDerivedFrom*—from PROV-O. Most of these attributes are optional and only 6 of them are required, these are: *@context*, *@type*, *identifier*, *name*, *author* and *datePublished*. Many of these attributes come from the JSON-LD metadata exported from the repositories we analyzed, however to increase the FAIRness of the metadata, we have added: *hasPart*, *conditionsOfAccess*, *wasDerivedFrom* and *isVersionOf*. The rationale behind the choice of these particular attributes has been provided in Section 4.3.

3.3.4 Additional FDOs

Each of the attributes mentioned in Section 3.3.3 requires an appropriate attribute definition. In this section, we present the set of these necessary attribute definitions, profiles and types defined within the FDO framework presented in Section 3.2. We also present the full profile that we used to represent the metadata of a record from a research data repository—invenio-record-profile. We named it this way, given the fact that in our research, we focus on this particular repository framework (InvenioRDM), however the name itself can be changed.

object-attribute-definition

The object-attribute-definition FDO has been depicted in Listing 3.14. This attribute definition is used to describe the *@context* attribute, which is the element that is used for integration with LOD vocabularies. It is defined as a JSON object, which allows for a lot of flexibility when choosing the right LOD annotations. The definition itself might strike as similar to the one of the *schema* attribute, however we distinguish between the two to make it easier to introduce any alterations, should they be needed in the future.

Listing 3.14: The object-attribute-definition FDO.

```
{
  "id": "object-attribute-definition",
  "fdoType": "CORDRA_PREFIX/FDO-attribute-definition-type",
  "fdoProfile": "CORDRA_PREFIX/FDO-content-profile",
  "schema": {"type": "object"}
}
```

text-attribute-definition

The text-attribute-definition FDO has been provided in Figure 3.15. This attribute definition is defined as a single *string* and is used to describe the following attributes:

- *@type*,
- *identifier*,
- *name*,
- *keywords*,
- *datePublished*,
- *dateModified*,
- *contentSize*,
- *version*,
- *license*,
- *description*,
- *url*,
- *conditionsOfAccess*,
- *wasDerivedFrom*,
- *isVersionOf*.

Listing 3.15: The text-attribute-definition FDO.

```
{
  "id": "text-attribute-definition",
  "fdoType": "CORDRA_PREFIX/FDO-attribute-definition-type",
  "fdoProfile": "CORDRA_PREFIX/FDO-content-profile",
  "schema": {"type": "string"}
}
```

contributors-attribute-definition

The contributors-attribute-definition FDO has been provided in Figure 3.16. This attribute definition is defined as an array of objects that contain the: *name*, *affiliation*, *@type* and *@id* properties. Additionally, the objects can also have the *givenName* and *familyName* properties. This FDO has been used to describe the *author* and *editor* attributes.

Listing 3.16: The contributors-attribute-definition FDO.

```
{
    "id": "contributors-attribute-definition",
    "fdoType": "CORDRA_PREFIX/FDO-attribute-definition-type",
    "fdoProfile": "CORDRA_PREFIX/FDO-content-profile",
    "schema": {
        "type": "array",
        "items": {
        "oneOf": [
            {
            "type": "object",
            "properties": {
                "name": {
                "type": "string"
                },
                "affiliation": {
                "type": "array",
                "items": {
                    "type": "object",
                    "properties": {
                    "@type": {
                        "type": "string"
                    },
                    "name": {
                        "type": "string"
                    }
                    },
                    "required": [
                    "@type",
                    "name"
                    ],
                    "additionalParameters": false
                }
                },
                "@id": {
                "type": "string"
                },
                "@type": {
                "type": "string"
                }
            },
            "required": [
                "name"
            ],
```

```
44                  "additionalProperties": false
45                  },
46                  {
47                  "type": "object",
48                  "properties": {
49                      "name": {
50                      "type": "string"
51                      },
52                      "givenName": {
53                      "type": "string"
54                      },
55                      "familyName": {
56                      "type": "string"
57                      },
58                      "affiliation": {
59                      "type": "array",
60                      "items": {
61                          "type": "object",
62                          "properties": {
63                          "@type": {
64                              "type": "string"
65                          },
66                          "name": {
67                              "type": "string"
68                          }
69                          },
70                          "required": [
71                          "@type",
72                          "name"
73                          ],
74                          "additionalParameters": false
75                      }
76                      },
77                      "@type": {
78                      "type": "string"
79                      },
80                      "@id": {
81                      "type": "string"
82                      }
83                  },
84                  "required": [
85                      "familyName"
86                  ],
87                  "additionalProperties": false
88                  }
89              ]
90              }
91          }
92      }
```

array-attribute-definition

The array-attribute-definition FDO has been presented in Figure 3.17. This attribute definition describes an array of any JSON objects. It has been used to describe the *hasPart* attribute.

Listing 3.17: The array-attribute-definition FDO.

```
1    {
2        "id": "array-attribute-definition",
3        "fdoType": "CORDRA_PREFIX/FDO-attribute-definition-type",
4        "fdoProfile": "CORDRA_PREFIX/FDO-content-profile",
5        "schema": {
6            "type": "array"
7        }
8    }
```

publisher-attribute-definition

The publisher-attribute-definition FDO has been presented in Figure 3.18. This attribute definition describes an object that has two properties: *@type* and *name*. It has been used to describe the *publisher* attribute.

Listing 3.18: The publisher-attribute-definition FDO.

```
{
    "id": "publisher-attribute-definition",
    "fdoType": "CORDRA_PREFIX/FDO-attribute-definition-type",
    "fdoProfile": "CORDRA_PREFIX/FDO-content-profile",
    "schema": {
      "type": "object",
      "properties": {
        "@type": {
          "type": "string"
        },
        "name": {
          "type": "string"
        }
      },
      "required": [
        "@type",
        "name"
      ],
      "additionalProperties": false
    }
  }
```

language-attribute-definition

The language-attribute-definition FDO has been presented in Figure 3.19. This attribute definition describes an object that has three properties: *alternateName*, *@type* and *name*. It has been used to describe the *inLanguage* attribute.

Listing 3.19: The language-attribute-definition FDO.

```
{
    "id": "language-attribute-definition",
    "fdoType": "CORDRA_PREFIX/FDO-attribute-definition-type",
    "fdoProfile": "CORDRA_PREFIX/FDO-content-profile",
    "schema": {
      "type": "object",
      "properties": {
        "alternateName": {
          "type": "string"
        },
        "@type": {
          "type": "string"
        },
        "name": {
          "type": "string"
        }
      },
      "required": [
        "name",
        "@type"
      ],
      "additionalProperties": false
    }
  }
```

citation-attribute-definition

The citation-attribute-definition FDO has been presented in Figure 3.20. This attribute definition describes an array of any JSON objects that have two properties: *@type* and *@id*. It has been used to describe the *citation* attribute.

Listing 3.20: The citation-attribute-definition FDO.

```
{
    "id": "citation-attribute-definition",
    "fdoType": "CORDRA_PREFIX/FDO-attribute-definition-type",
    "fdoProfile": "CORDRA_PREFIX/FDO-content-profile",
    "schema": {
      "type": "array",
      "items": {
        "type": "object",
        "properties": {
          "@type": {
            "type": "string"
          },
          "@id": {
            "type": "string"
          }
        },
        "required": [
          "@type",
          "@id"
        ],
        "additionalProperties": false
      }
    }
  }
```

distribution-attribute-definition

The distribution-attribute-definition FDO has been presented in Figure 3.21. This attribute definition describes an array of any JSON objects that have three properties: *url*, *fileFormat* and *fileSize*. It has been used to describe the *distribution* attribute.

Listing 3.21: The distribution-attribute-definition FDO.

```
{
    "id": "distribution-attribute-definition",
    "fdoType": "CORDRA_PREFIX/FDO-attribute-definition-type",
    "fdoProfile": "CORDRA_PREFIX/FDO-content-profile",
    "schema": {
      "type": "array",
      "items": {
        "type": "object",
        "properties": {
          "url": {
            "type": "string"
          },
          "fileFormat": {
            "type": "string"
          },
          "fileSize": {
            "type": "number"
          }
        },
        "required": [
          "url",
          "fileFormat",
          "fileSize"
        ],
        "additionalProperties": false
      }
    }
  }
```

invenio-record-type

The invenio-record-type FDO has been presented in Figure 3.22. It provides a human-readable description that explains what FDOs of this type represent. It has been used to describe every FDO that represents metadata of a certain record from InvenioRDM.

Listing 3.22: The invenio-record-type FDO.

```
1    {
2        "id": "invenio-record-type",
3        "fdoType": "CORDRA_PREFIX/invenio-record-type",
4        "fdoProfile": "CORDRA_PREFIX/FDO-type-profile",
5        "description": "This is a record imported from an invenioRDM instance
             "
6      }
```

invenio-record-profile

The invenio-record-profile FDO has been presented in Figure 3.23. This profile is used to provide a machine-actionable description to every FDO that represents metadata of a certain record from a research data repository. It describes object's attributes using the FDOs from previous sections and defines the attributes mentioned in Section 3.3.3.

Listing 3.23: The invenio-record-profile FDO.

```
{
    "id": "invenio-record-profile",
    "fdoType": "CORDRA_PREFIX/FDO-profile-type",
    "fdoProfile": "CORDRA_PREFIX/FDO-content-profile",
    "schema": {
      "allOf": [
        {
          "$ref": "FDO-configuration"
        },
        {
          "type": "object",
          "properties": {
            "@context": "FDO_REF:CORDRA_PREFIX/object-attribute-definition",
            "@type": "FDO_REF:CORDRA_PREFIX/text-attribute-definition",
            "identifier": "FDO_REF:CORDRA_PREFIX/text-attribute-definition",
            "name": "FDO_REF:CORDRA_PREFIX/text-attribute-definition",
            "author": "FDO_REF:CORDRA_PREFIX/contributors-attribute-definition",
            "editor": "FDO_REF:CORDRA_PREFIX/contributors-attribute-definition",
            "hasPart": "FDO_REF:CORDRA_PREFIX/array-attribute-definition",
            "publisher": "FDO_REF:CORDRA_PREFIX/publisher-attribute-definition",
            "keywords": "FDO_REF:CORDRA_PREFIX/text-attribute-definition",
            "datePublished": "FDO_REF:CORDRA_PREFIX/text-attribute-definition",
            "dateModified": "FDO_REF:CORDRA_PREFIX/text-attribute-definition",
            "inLanguage": "FDO_REF:CORDRA_PREFIX/language-attribute-definition",
            "contentSize": "FDO_REF:CORDRA_PREFIX/text-attribute-definition",
            "version": "FDO_REF:CORDRA_PREFIX/text-attribute-definition",
            "license": "FDO_REF:CORDRA_PREFIX/text-attribute-definition",
            "description": "FDO_REF:CORDRA_PREFIX/text-attribute-definition",
            "citation": "FDO_REF:CORDRA_PREFIX/citation-attribute-definition",
            "url": "FDO_REF:CORDRA_PREFIX/text-attribute-definition",
            "distribution": "FDO_REF:CORDRA_PREFIX/distribution-attribute-definition",
```

```
32                "conditionsOfAccess": "FDO_REF:CORDRA_PREFIX/text-attribute-
                      definition",
33                "wasDerivedFrom": "FDO_REF:CORDRA_PREFIX/text-attribute-
                      definition",
34                "isVersionOf": "FDO_REF:CORDRA_PREFIX/text-attribute-
                      definition"
35              },
36              "required": [
37                "@context",
38                "@type",
39                "identifier",
40                "name",
41                "author",
42                "datePublished"
43              ]
44            }
45          ],
46          "unevaluatedProperties": false
47        }
48      }
```

3.4 Migration procedure

One of the primary goals of this thesis is to present a conceptual workflow that describes how to migrate the metadata from an existing data repository to a chosen FAIR Digital Objects (FDOs) backend. For this purpose, we created a procedure called migration that executes the following steps:

1. create the necessary core elements of the FDOs framework in the FDO backend if it has not been done already;
2. create the additional FDOs from the *Research Metadata Extension* necessary to accurately describe repository records' metadata (profiles, types and attribute definitions);
3. migrate the metadata from the repository to the FDO backend.

In this procedure, we identify three services that take part in it:

- Repository—the research data repository that contains the records,
- Migration service—an entity that executes the migration procedure and
- FDO backend—the service where the FDOs are stored.

We present how these services interact with each other in Figure 3.4.

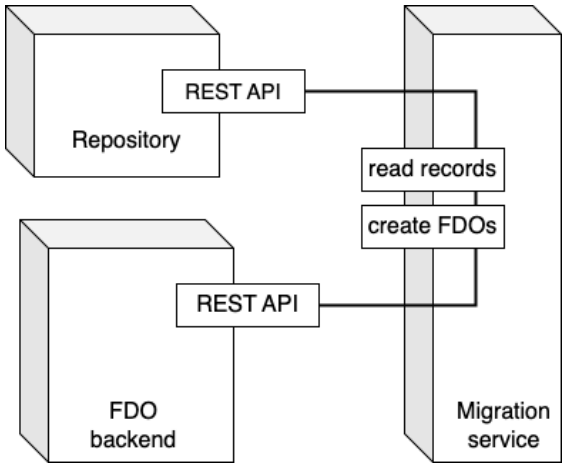

Figure 3.4: Diagram of presenting how the services in the migration procedure interact with each other.

The simplified flow of this procedure has been depicted in the diagram 3.5. Whenever an administrator starts the migration procedure, it performs necessary actions using REST

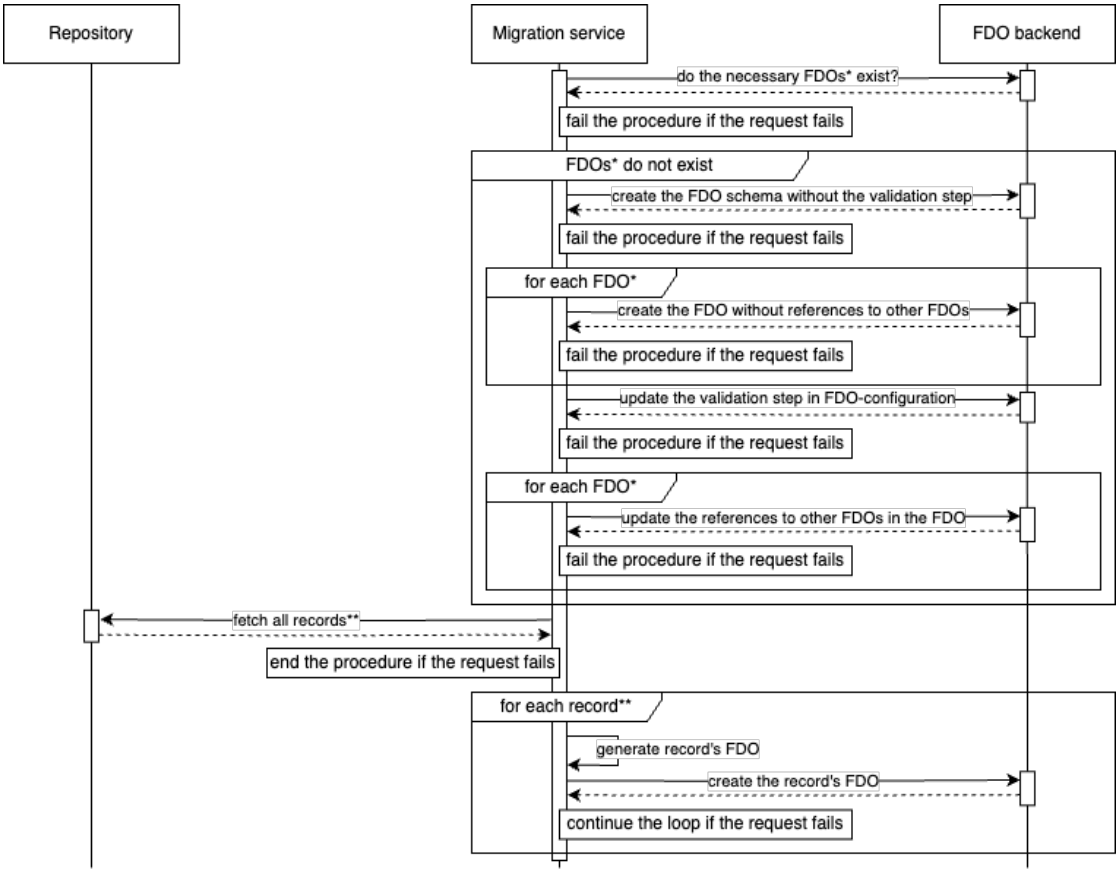

Figure 3.5: Diagram of the migration procedure described in points 1 to 3. (*)—by FDOs, we mean the FDO-configuration schema, 6 core FDOs and the rest of other necessary profiles, types and attribute definitions created for the use case studied in this thesis. (**)—do not fetch records that are *metadata-only*.

APIs of the targeted repository and the FDO backend. Firstly, it checks which objects are already defined in the FDO backend. It is necessary that the following objects exist:

- the FDO-configuration schema,
- the core FDOs from Section 3.2 and
- FDOs from the *Research Metadata Extension* necessary to create objects that follow the *invenio-record-profile* described in Section 3.3.3.

If one of them is missing, the procedure recreates them to make sure that the FDOs representing the metadata from the targeted repository can be added to the FDO backend. To create these required objects, the procedure has to accommodate for circular dependencies in the design. Initially, the FDO-configuration is created without the extended validation

mechanism that validates objects based on their profiles. Subsequently, the necessary FDOs are created, however without the references in their *fdoType* and *fdoProfile* attributes. This is necessary because some FDOs contain references to themselves, such as the FDO-content-profile and FDO-attribute-definition-type. Afterwards, the validation mechanism in the FDO-configuration is uploaded to the FDO backend and the FDOs are updated with the appropriate references. However, during this update, the validation procedure is carried out, which ensures that the core building blocks are valid in terms of this framework.

After it has been made sure that the necessary FDOs exist in the FDO backend, the migration itself takes place. Firstly, the list of records from the targeted repository along with their JSON-LD metadata and information about related files is acquired. Based on this data, the FDOs are created and uploaded to the FDO backend. Most attributes in the new metadata representation come from the JSON-LD metadata exported from the repository. However, the tool adds certain fields, especially *prov:wasDerivedFrom* and *schema:distribution*, which in turn, increases the FAIRness score (see Section 4.3).

This procedure is intended to be used by an administrator of an already existing research data repository, who intends to elevate the FAIRness of records in that repository. This person can use this procedure to migrate the metadata of records from their repository to our novel FDO framework. It is fairly straightforward and should not require a significant amount of manual labor. We provide more implementation details and information about how our solution influences the said FAIRness in Chapter 4.

3.5 Discussion

In this chapter, we presented the novel FDO framework that we developed along with all the FAIR Digital Objects that we found necessary to model the metadata of research data repository records in this framework—the *Research Metadata Extension*. We also presented how the objects are integrated with LOD vocabularies, which increases the FAIRness of these objects. At the end of this chapter, we also presented the migration procedure.

Therefore, in this chapter, we provided our answer to the first research question from Section 1.1.1. We present the appropriate FDO profiles and attributes that can be used to represent records in a research data repository as FDOs. These are the core FDOs from Section 3.2.1:

- FDO-content-profile,
- FDO-type-profile,
- FDO-attribute-definition-type,
- FDO-profile-type,

- FDO-schema-attribute-definition and
- FDO-description-attribute-definition

along with the Cordra Schema FDO-configuration and additional FDOs from the *Research Metadata Extension* from Section 3.3.4.

CHAPTER 4

Evaluation

In this chapter, we evaluate the effectiveness of the migration procedure from Section 3.4 using a tool that migrates the metadata from a research data repository of our preference (an instance of the InvenioRDM project) to the chosen FAIR Digital Object (FDO) backend—Cordra. We also provide the conceptual architecture of a research data repository based on InvenioRDM that represents datasets as FDOs, which is a direct answer to the second research question from Section 1.1.2.

We also evaluate the FAIRness of the metadata using the automated tool FUJI[1] and we attempt to increase this score. Initially, we perform tests based on the data from the test instance of the TU Wien Research Data repository to identify areas where an improvement of the FAIRness is possible. Secondly, we identify which attributes positively influence the score by enriching the metadata with new fields and testing how the FAIRness score fluctuates. Section 4.2 and Section 4.3 describe the results of these tests and which design decisions have been made regarding the choice of additional attributes.

This section contains, thus, also the answer to the third research question from Section 1.1.3—Section 4.4 neatly summarizes the advantages of adopting the proposed solution with regard to the FAIR principles.

4.1 Migration Assistant architecture

To evaluate the effectiveness of the migration procedure from Section 3.4, we have created a tool that migrates the metadata from a research data repository of our preference to the chosen FDO backend—Cordra. This tool works with repositories based on the InvenioRDM project, since we focus on such repositories (as mentioned in Section 2.6). It

[1]We used the FUJI version found in: `https://github.com/pangaea-data-publisher/fuji.git` and worked with the version available at the commit: 42472ac

is a Python Flask[2] web application called *Migration Assistant*[3] that works as a translation layer between the two services. Its main functionalities are to:

1. carry out the migration procedure from Section 3.4 and
2. expose an endpoint for FAIRness assessment of the newly created FDOs suitable for FUJI.

The code has been published in a public repository—please see the appendix A for more details.

The architecture of the *Migration Assistant* application together with the instances of InvenioRDM and Cordra represents an example of how a data repository system that exports the metadata of its records as FDOs could be structured. It connects the instances of InvenioRDM and Cordra using their respective REST APIs and it exposes Cordra's DOIP endpoint for testing using FUJI. A high-level overview has been depicted in Figure 4.1.

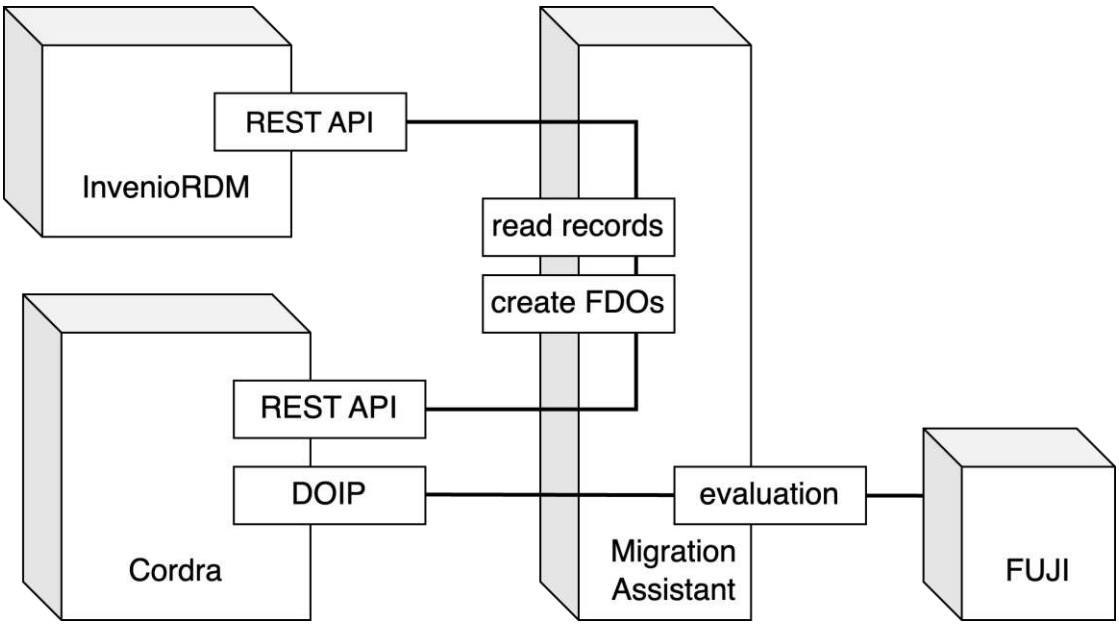

Figure 4.1: Diagram of the Migration Assistant architecture

Moreover, given the fact that the chosen assessment tool—FUJI evaluates not only the quality of the metadata but also its availability, it has to acquire the metadata in a specific way. It is important that the tool is able to access the metadata from a *script* tag found in a webpage and using the content negotiation mechanism. Otherwise, the score is lowered due to reasons unrelated with the FAIRness of the metadata itself[4]. This, in

[2] `https://flask.palletsprojects.com/en/stable/`
[3] `https://github.com/Brotholomew/migration_assistant`
[4] Testing has shown that the FsF-F4-01M-1 FUJI test fails only due to the accessibility of the metadata.

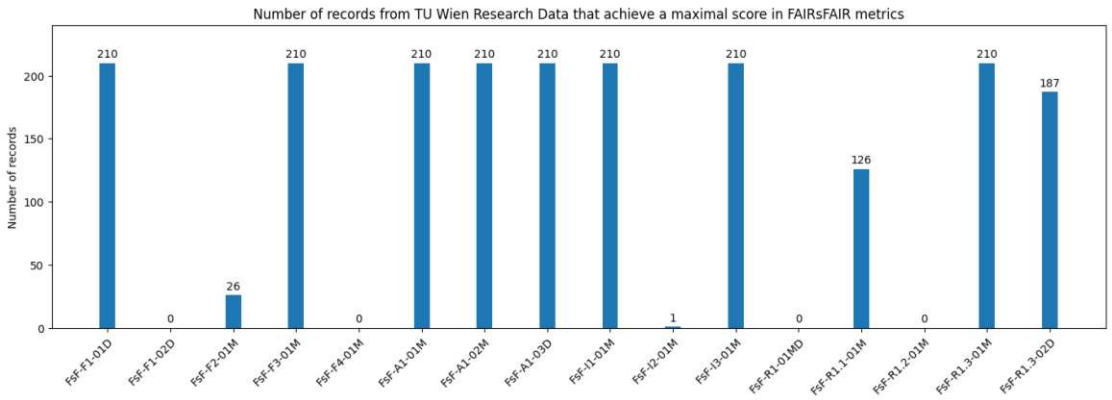

Figure 4.2: Number of records from TU Wien Research Data that achieve a maximal score in each of the FAIRsFAIR metrics.

turn, makes it difficult to compare the FAIRness of metadata from InvenioRDM with the FAIRness of FDOs exported from Cordra. Therefore, the *Migration Assistant* tool also exposes an endpoint */metadata/<record-id>* to allow for a straightforward evaluation process using FUJI. When a GET request reaches this endpoint, the tool exports the appropriate FDO from Cordra in the JSON format and returns an HTML document that contains a *script* tag with the metadata. However, because the FDO contains JSON-LD keywords, the tag contains the *type="application/ld+json"* annotation. The endpoint also returns the exported FDO during the content negotiation part of the assessment when the accepted mime type of the request is set to *application/ld+json*.

The usage of this tool, from a system administrator's perspective, is relatively straightforward. All that this person has to do is to set up the application and provide credentials to their instances of Cordra and InvenioRDM. Afterwards, the migration can be carried out with a click of a single button. Therefore, a low amount of manual labor is required to adopt this solution. Moreover, after the migration process, the InvenioRDM records remain intact, which makes the whole procedure risk-free.

4.2 FAIRness of metadata from InvenioRDM

The FUJI tool, apart from assessing the combined FAIRness score (as depicted in Figure 2.3) also measures how well a given FAIRsFAIR metric has been satisfied. It allocates a certain amount of points for each metric. This is summed up at the end to create the final FAIRness score. Each of the 16 FAIRsFAIR metrics[5] that FUJI uses are evaluated within 47 tests. The overall maximal score is 24 points (FAIRness score of 100%).

[5]It has been previously mentioned that there are 17 FAIRsFAIR metrics, however FUJI does not check the conditions of *FsF-A2-01M*—the metadata should be persisted, even if the data is no longer available.

We have evaluated 210 records from TU Wien Research Data and measured how many of them receive the maximal amount of points in each FAIRsFAIR metric. The results have been depicted in Figure 4.2. All records receive the maximal amount of points in the metrics[6]:

- *FsF-F1-01D*—a globally unique PID is assigned to both the metadata and the data;
- *FsF-F3-01M*—metadata includes an identifier of the data it describes;
- *FsF-A1-01M*—metadata contains access level and access conditions of the data;
- *FsF-A1-02M*—metadata is accessible through a standardized communication protocol;
- *FsF-A1-03D*—data is accessible through a standardized communication protocol;
- *FsF-I1-01M*—metadata is represented using a formal knowledge representation language;
- *FsF-I3-01M*—metadata includes links between the data and its related entities;
- *FsF-R1.3-01M*—metadata follows a standard recommended by the target research community of the data.

Given the fact that there is no need for improvement in terms of the aforementioned components of the FAIRness score, we shift our focus to the other ones. Moreover, the *FsF-F1-02D* metric is unsatisfied for reasons unrelated with metadata structure. It requires that the metadata is assigned a PID, however FUJI demands more than this. The tool also checks if the PID is resolvable and, if so, if it resolves to the same origin as the url provided for assessment. However, since we are evaluating the test instance of TU Wien Research Data, the origin differs and FUJI does not assign any points in this metric.[7]

Similarly, no record got a perfect score in the *FsF-F4-01M* metric, which requires that the metadata is offered in such a way that it can be retrieved programmatically. Even though the metadata can be retrieved programmatically, the metric is not fully satisfied because it is not found through registries considered by the assessment service, such as e.g.: DataCite, which is a problem unrelated to metadata structure.

Furthermore, the data posted in many knowledge bases is oftentimes incomplete [MTDF17, DNPR18]. The same problem regards the records published in TU Wien Research Data repository. Some records have been published without filling out all fields in the InvenioRDM UI, and therefore the metrics *FsF-F2-01M*, *FsF-R1.1-01M* and *FsF-R1.3-02D* are not satisfied. For the first one, it is crucial that the core descriptive metadata

[6]The definitions of the metrics are taken from [DHM+22]

[7]FUJI returns this warning in the logs: *PID syntax is OK but the PID seems to resolve to a different entity, will not use this PID for content negotiation.*

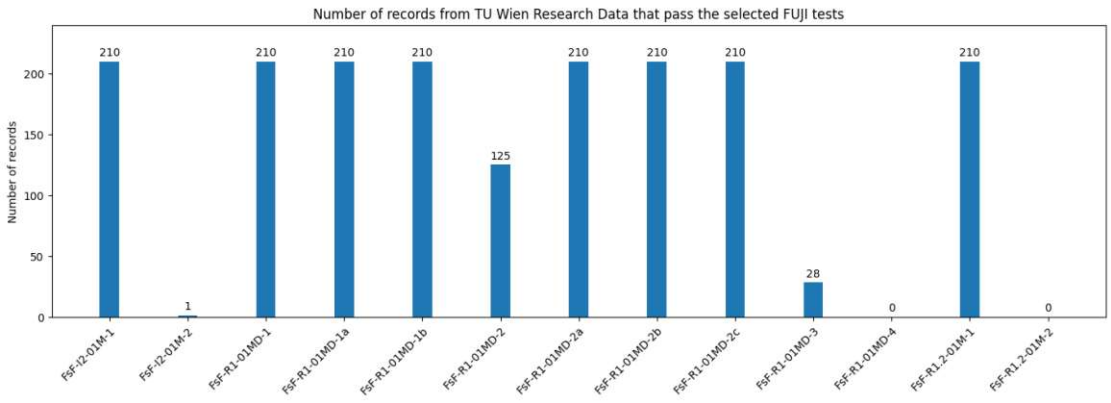

Figure 4.3: Number of records from TU Wien Research Data that achieve a maximal score in the selected FUJI tests

elements exist, such as: name, creator, title, summary, keywords, etc. The second metric requires that some licensing information is given in the metadata. Lastly, the *FsF-R1.3-02D* metric demands that the data described in the metadata is available in a file format recommended by the target research community. It is possible to find an example[8] that invalidates all three mentioned metrics due to incompleteness of its metadata. It lacks the core descriptive attribute: keywords, there is no licensing information provided and FUJI is unable to check if the last metric is satisfied. The reason that not all of the records achieve a full score in these three metrics is, thus, unrelated to metadata structure, it is caused by its incompleteness.

For the above reasons, we exclude the mentioned metrics from further analyses and focus on the remaining ones:

1. *FsF-I2-01M*—metadata uses semantic resources;
2. *FsF-R1-01MD*—metadata specifies the content of the data;
3. *FsF-R1.2-01M*—metadata includes provenance information about data creation or generation.

To gain a deeper insight on how the scores could be improved, we analyzed the results of FUJI tests related to these three metrics. As depicted in Figure 4.3, there is room for improvement in terms of the following tests:

1. *FsF-I2-01M-2*—namespaces of known semantic resources can be identified in metadata;
2. *FsF-R1-01MD-2*—verifiable data descriptors are specified in metadata;

[8] `https://test.researchdata.tuwien.at/records/ndwv5-31f90`

3. *FsF-R1-01MD-3*—data content matches file type and size or protocol specified in metadata;

4. *FsF-R1-01MD-4*—data content matches measured variables or observation types specified in metadata;

5. *FsF-R1.2-01M-2*—metadata contains provenance information using formal provenance ontologies (PROV-O).

We present some chosen records that failed the aforementioned FUJI tests along with appropriate FUJI test logs in the following tables. The Table 4.1 contains records that failed the *FsF-I2-01M-2* test. Table 4.2 covers tests: *FsF-R1-01MD-2* and *FsF-R1-01MD-3*. Table 4.3 contains records that failed the *FsF-R1.2-01M-2* test.

Record	URL	FUJI test log
yvh49-kvd12	`https://test.researchdata.tuwien.ac.at/records/yvh49-kvd12`	WARNING: NO known vocabulary namespace URI is found which is listed in the LOD registry
3tf7q-e9t81	`https://test.researchdata.tuwien.ac.at/records/3tf7q-e9t81`	WARNING: NO known vocabulary namespace URI is found which is listed in the LOD registry
jcpn6-vnm19	`https://test.researchdata.tuwien.ac.at/records/jcpn6-vnm19`	WARNING: NO known vocabulary namespace URI is found which is listed in the LOD registry
05mkq-2xz41	`https://test.researchdata.tuwien.ac.at/records/05mkq-2xz41`	WARNING: NO known vocabulary namespace URI is found which is listed in the LOD registry
26sb2-d7635	`https://test.researchdata.tuwien.ac.at/records/26sb2-d7635`	WARNING: NO known vocabulary namespace URI is found which is listed in the LOD registry

Table 4.1: Exemplary records that failed the *FsF-I2-01M-2* test (5 out of 209 records in total).

Record	URL	FUJI test log
qxbm7-1md23	https://test.researchdata.tuwien.ac.at/records/qxbm7-1md23	INFO: NO info about data service endpoint available in given metadata for -: https://test.researchdata.tuwien.at/api/records/qxbm7-1md23/files/username.csv/content WARNING: Could not verify content type from downloaded file -: (expected: text/csv, found: via tika ['text/plain'] or via header text/plain)
dc4zh-9ce78	https://test.researchdata.tuwien.ac.at/records/dc4zh-9ce78	INFO: NO info about data service endpoint available in given metadata for -: https://test.researchdata.tuwien.at/records/dc4zh-9ce78/files/colive.0066_20200611134530_1_m4a_0.wav WARNING: Could not verify content type from downloaded file -: (expected: audio/x-wav, found: via tika ['audio/vnd.wave'] or via header application/octet-stream)
nca1z-y3d12	https://test.researchdata.tuwien.ac.at/records/nca1z-y3d12	INFO: NO info about data service endpoint available in given metadata for -: https://test.researchdata.tuwien.at/records/nca1z-y3d12/files/colive.0044_20200518133554_1_m4a_1.wav WARNING: Could not verify content type from downloaded file -: (expected: audio/x-wav, found: via tika ['audio/vnd.wave'] or via header application/octet-stream)
w37m8-dx896	https://test.researchdata.tuwien.ac.at/records/w37m8-dx896	INFO: NO info about data service endpoint available in given metadata for -: https://test.researchdata.tuwien.at/records/w37m8-dx896/files/colive.0044_20200518133554_1_m4a_1.wav WARNING: Could not verify content type from downloaded file -: (expected: audio/x-wav, found: via tika ['audio/vnd.wave'] or via header application/octet-stream)

0f99d-w2e63	https://test.researchdata.tuwien.ac.at/records/0f99d-w2e63	INFO: NO info about data service endpoint available in given metadata for -: https://test.researchdata.tuwien.at/api/records/0f99d-w2e63/files/username.csv/content WARNING: Could not verify content type from downloaded file -: (expected: text/csv, found: via tika ['text/plain'] or via header text/plain)
rn3hz-khe04	https://test.researchdata.tuwien.ac.at/records/rn3hz-khe04	INFO: NO info about data service endpoint available in given metadata for -: https://test.researchdata.tuwien.at/records/rn3hz-khe04/files/colive.0066_20200611134530_2_m4a_0.wav WARNING: Could not verify content type from downloaded file -: (expected: audio/x-wav, found: via tika ['audio/vnd.wave'] or via header application/octet-stream)

Table 4.2: Exemplary records that failed the *FsF-R1-01MD-2* and *FsF-R1-01MD-3* tests (6 out of 85 records in total).

Record	URL	FUJI test log
yvh49-kvd12	https://test.researchdata.tuwien.ac.at/records/yvh49-kvd12	WARNING: Formal provenance metadata is unavailable
3tf7q-e9t81	https://test.researchdata.tuwien.ac.at/records/3tf7q-e9t81	WARNING: Formal provenance metadata is unavailable
jcpn6-vnm19	https://test.researchdata.tuwien.ac.at/records/jcpn6-vnm19	WARNING: Formal provenance metadata is unavailable
05mkq-2xz41	https://test.researchdata.tuwien.ac.at/records/05mkq-2xz41	WARNING: Formal provenance metadata is unavailable

26sb2-d7635	`https://test.researchdata.tuwien.ac.at/records/26sb2-d7635`	WARNING: Formal provenance metadata is unavailable

Table 4.3: Exemplary records that failed the *FsF-R1.2-01M-2* test (5 out of 210 records in total).

4.3 FAIRness of the new metadata format

To establish which attributes should be included in the new metadata format, we analyzed the files exported by InvenioRDM during the content negotiation part of FUJI assessment. First of all, it exports the following schema.org annotations in the JSON-LD format of the metadata: *identifier*, *name*, *author*, *editor*, *publisher*, *keywords*, *datePublished*, *dateModified*, *inLanguage*, *contentSize*, *version*, *license*, *description*, *citation* and *url*. If a given record is created as a dataset, the JSON-LD graph also contains the *distribution* property, which contains information about the published files. The *distribution* attribute, however, is missing in case of other types of records. Secondly, FUJI also acquires information about access conditions to the data (*schema.org:conditionsOfAccess*), however this is not in the JSON-LD metadata—InvenioRDM exposes this information using different metadata formats. Similarly, attributes from DCMI Metadata Terms: *hasPart* and *isVersionOf* are also recognized by FUJI, however they are not included in the JSON-LD metadata.

To make sure that the FAIRness score of the new metadata format is not worse than the score of metadata from InvenioRDM, we added the aforementioned properties to our solution. This means that we created appropriate attribute definitions and included them in the *schema* attribute of the *invenio-record-profile* FDO (as mentioned in Section 3.3.3). This includes the attributes stemming from schema.org and also DCMI Metadata Terms, therefore, these two vocabularies have been included in the *@context* of every FDO that represents metadata of InvenioRDM records.

To improve the FAIRness score, we started with analyzing the results of the *FsF-R1-01MD-2* (2) and *FsF-R1-01MD-3* (3) FUJI tests. It turned out that the lack of the *distribution* attribute in the metadata for types other than *Dataset* decreases the score of the first test. The latter test fails because, for some records, FUJI either cannot access the information about the file size of the data attached to a record[9], or it finds a wrong file size[10]. We attempt to mitigate these issues by ensuring that every record which has some files attached to it has the *distribution* attribute in its metadata. Moreover,

[9]This is the case, for example, for this record: `https://test.researchdata.tuwien.ac.at/records/26sb2-d7635`

[10]This happens, for example, for this record: `https://test.researchdata.tuwien.at/records/jcpn6-vnm19`

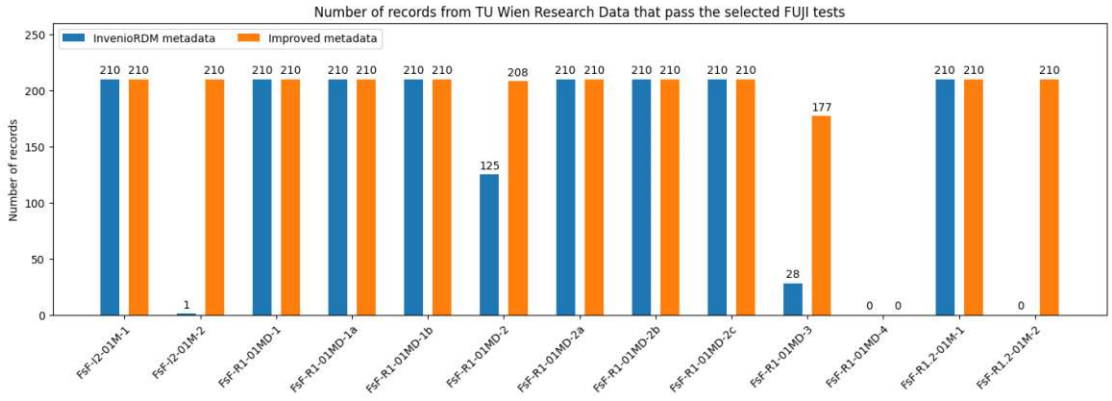

Figure 4.4: Number of records from TU Wien Research Data that achieve a maximal score in the selected FUJI tests. A juxtaposition of results between the metadata from InvenioRDM and the newly created format of metadata.

the file size (in bytes) and file mime type information is taken from the InvenioRDM *api/records/<record-id>/files* REST API endpoint, which ensures that it is correct.

During the test *FsF-I2-01M-2* (1), FUJI tries to establish if the metadata also contains other taxonomies apart from the commonly used ones. We enrich the metadata by adding a property from the PROV-O ontology: *wasDerivedFrom*. The value of this new attribute contains a URL to the record that the metadata describes. This serves as an indication of provenance regarding the metadata object itself—it has been derived from the information about this particular record that it references in this attribute. This should also positively influence the result of the *FsF-R1.2-01M-2* (5) test.

We did not find any use cases in the TU Wien Research Data repository that would have information about measured variables in their metadata. Therefore, we do not propose any changes or new attributes that would increase the score in terms of the *FsF-R1-01MD-4* test (4).

All in all, the proposed metadata format achieves a higher FAIRness score. As depicted in Figure 4.4, all records have passed the tests: *FsF-I2-01M-2* (1) and *FsF-R1.2-01M-2* (5). The number of records that passed *FsF-R1-01MD-2* (2) is 208—the two records that failed the test contain empty files[11], which apparently is not accepted by FUJI. Furthermore, the test *FsF-R1-01MD-3* (3) has been passed by 177 records—33 records failed this test due to file type inconsistencies. FUJI failed to recognize the following mime types extracted from InvenioRDM *api/records/<record-id>/files* REST API endpoint:

- *audio/x-wav*,
- *application/pdf*,

[11]These are: `https://test.researchdata.tuwien.ac.at/records/ecbjk-7ka62` and `https://test.researchdata.tuwien.ac.at/records/apwsf-ejr45`

- *application/x-netcdf*,
- *text/csv* and
- *text/x-python*.

We present the records that failed both the *FsF-R1-01MD-2* and *FsF-R1-01MD-3* FUJI tests along with appropriate FUJI test logs in Table 4.4.

Record	URL	FUJI test log
jcpn6-vnm19	https://test.researchdata.tuwien.ac.at/records/jcpn6-vnm19	WARNING: Could not verify content type from downloaded file -: (expected: audio/x-wav, found: via tika ['audio/vnd.wave'] or via header application/octet-stream)
05mkq-2xz41	https://test.researchdata.tuwien.ac.at/records/05mkq-2xz41	WARNING: Could not verify content type from downloaded file -: (expected: application/pdf, found: via tika ['text/plain'] or via header application/octet-stream)
dc4zh-9ce78	https://test.researchdata.tuwien.ac.at/records/dc4zh-9ce78	WARNING: Could not verify content type from downloaded file -: (expected: audio/x-wav, found: via tika ['audio/vnd.wave'] or via header application/octet-stream)
pqb7y-mtf49	https://test.researchdata.tuwien.ac.at/records/pqb7y-mtf49	WARNING: Could not verify content type from downloaded file -: (expected: audio/x-wav, found: via tika ['audio/vnd.wave'] or via header application/octet-stream)
t3zj3-bns11	https://test.researchdata.tuwien.ac.at/records/t3zj3-bns11	WARNING: Could not verify content type from downloaded file -: (expected: audio/x-wav, found: via tika ['audio/vnd.wave'] or via header application/octet-stream)

nca1z-y3d12	https://test.researchdata.tuwien.ac.at/records/nca1z-y3d12	WARNING: Could not verify content type from downloaded file -: (expected: audio/x-wav, found: via tika ['audio/vnd.wave'] or via header application/octet-stream)
kqrnz-3nh51	https://test.researchdata.tuwien.ac.at/records/kqrnz-3nh51	WARNING: Could not verify content type from downloaded file -: (expected: application/x-netcdf, found: via tika ['application/x-hdf', 'application/hdf'] or via header application/octet-stream)
w37m8-dx896	https://test.researchdata.tuwien.ac.at/records/w37m8-dx896	WARNING: Could not verify content type from downloaded file -: (expected: audio/x-wav, found: via tika ['audio/vnd.wave'] or via header application/octet-stream)
30vm3-f3e20	https://test.researchdata.tuwien.ac.at/records/30vm3-f3e20	WARNING: Could not verify content type from downloaded file -: (expected: text/csv, found: via tika ['text/plain'] or via header text/plain)
eym2v-nwy33	https://test.researchdata.tuwien.ac.at/records/eym2v-nwy33	WARNING: Could not verify content type from downloaded file -: (expected: audio/x-wav, found: via tika ['audio/vnd.wave'] or via header application/octet-stream)
kfzkm-fjb47	https://test.researchdata.tuwien.ac.at/records/kfzkm-fjb47	WARNING: Could not verify content type from downloaded file -: (expected: audio/x-wav, found: via tika ['audio/vnd.wave'] or via header application/octet-stream)
rn3hz-khe04	https://test.researchdata.tuwien.ac.at/records/rn3hz-khe04	WARNING: Could not verify content type from downloaded file -: (expected: audio/x-wav, found: via tika ['audio/vnd.wave'] or via header application/octet-stream)

8jh4p-fd291	https://test.researchdata.tuwien.ac.at/records/8jh4p-fd291	WARNING: Could not verify content type from downloaded file -: (expected: audio/x-wav, found: via tika ['audio/vnd.wave'] or via header application/octet-stream)
ch8jr-03d86	https://test.researchdata.tuwien.ac.at/records/ch8jr-03d86	WARNING: Could not verify content type from downloaded file -: (expected: audio/x-wav, found: via tika ['audio/vnd.wave'] or via header application/octet-stream)
hg60g-1hg73	https://test.researchdata.tuwien.ac.at/records/hg60g-1hg73	WARNING: Could not verify content type from downloaded file -: (expected: audio/x-wav, found: via tika ['audio/vnd.wave'] or via header application/octet-stream)
a13x6-2wz51	https://test.researchdata.tuwien.ac.at/records/a13x6-2wz51	WARNING: Could not verify content type from downloaded file -: (expected: audio/x-wav, found: via tika ['audio/vnd.wave'] or via header application/octet-stream)
xycj0-var90	https://test.researchdata.tuwien.ac.at/records/xycj0-var90	WARNING: Could not verify content type from downloaded file -: (expected: application/x-netcdf, found: via tika ['application/x-hdf', 'application/hdf'] or via header application/octet-stream)
tp5km-6ev59	https://test.researchdata.tuwien.ac.at/records/tp5km-6ev59	WARNING: Could not verify content type from downloaded file -: (expected: application/x-netcdf, found: via tika ['application/x-hdf', 'application/hdf'] or via header application/octet-stream)
tzf66-yag85	https://test.researchdata.tuwien.ac.at/records/tzf66-yag85	WARNING: Could not verify content type from downloaded file -: (expected: text/csv, found: via tika ['text/plain'] or via header text/plain)

22rr8-4d542	https://test.researchdata.tuwien.ac.at/records/22rr8-4d542	WARNING: Could not verify content type from downloaded file -: (expected: audio/x-wav, found: via tika ['audio/vnd.wave'] or via header application/octet-stream)
xksnz-cfg98	https://test.researchdata.tuwien.ac.at/records/xksnz-cfg98	WARNING: Could not verify content type from downloaded file -: (expected: audio/x-wav, found: via tika ['audio/vnd.wave'] or via header application/octet-stream)
vqpbr-5b889	https://test.researchdata.tuwien.ac.at/records/vqpbr-5b889	WARNING: Could not verify content type from downloaded file -: (expected: audio/x-wav, found: via tika ['audio/vnd.wave'] or via header application/octet-stream)
345zp-acm54	https://test.researchdata.tuwien.ac.at/records/345zp-acm54	WARNING: Could not verify content type from downloaded file -: (expected: application/x-netcdf, found: via tika ['application/x-hdf', 'application/hdf'] or via header application/octet-stream)
7qxy8-xk520	https://test.researchdata.tuwien.ac.at/records/7qxy8-xk520	WARNING: Could not verify content type from downloaded file -: (expected: text/x-python, found: via tika ['application/x-sh', 'application/sh'] or via header application/octet-stream)
6037v-32289	https://test.researchdata.tuwien.ac.at/records/6037v-32289	WARNING: Could not verify content type from downloaded file -: (expected: application/x-netcdf, found: via tika ['application/x-hdf', 'application/hdf'] or via header application/octet-stream)
0xf26-8ae60	https://test.researchdata.tuwien.ac.at/records/0xf26-8ae60	WARNING: Could not verify content type from downloaded file -: (expected: audio/x-wav, found: via tika ['audio/vnd.wave'] or via header application/octet-stream)

ntvp4-wxb61	https://test.researchdata.tuwien.ac.at/records/ntvp4-wxb61	WARNING: Could not verify content type from downloaded file -: (expected: audio/x-wav, found: via tika ['audio/vnd.wave'] or via header application/octet-stream)
rv5hk-fjh93	https://test.researchdata.tuwien.ac.at/records/rv5hk-fjh93	WARNING: Could not verify content type from downloaded file -: (expected: application/x-netcdf, found: via tika ['application/x-hdf', 'application/hdf'] or via header application/octet-stream)
ecbjk-7ka62	https://test.researchdata.tuwien.ac.at/records/ecbjk-7ka62	INFO: NO info about data service endpoint available in given metadata for -: https://test.researchdata.tuwien.at/api/records/ecbjk-7ka62/files/empty/content
pvs08-54b28	https://test.researchdata.tuwien.ac.at/records/pvs08-54b28	WARNING: Could not verify content type from downloaded file -: (expected: text/csv, found: via tika ['text/plain'] or via header text/plain)
j81zs-ejc61	https://test.researchdata.tuwien.ac.at/records/j81zs-ejc61	WARNING: Could not verify content type from downloaded file -: (expected: application/json, found: via tika ['text/plain'] or via header text/plain)
apwsf-ejr45	https://test.researchdata.tuwien.ac.at/records/apwsf-ejr45	INFO: NO info about data service endpoint available in given metadata for -: https://test.researchdata.tuwien.at/api/records/apwsf-ejr45/files/airdata_model.md5/content
wh7vb-f2a38	https://test.researchdata.tuwien.ac.at/records/wh7vb-f2a38	WARNING: Could not verify content type from downloaded file -: (expected: audio/x-wav, found: via tika ['audio/vnd.wave'] or via header application/octet-stream)

Table 4.4: All these records failed the *FsF-R1-01MD-3* test. Additionally, *apwsf-ejr45* and *ecbjk-7ka62* also failed *FsF-R1-01MD-2*.

Please refer to appendix B for a more detailed report about the aforementioned FUJI tests: *FsF-R1-01MD-2* (2), *FsF-R1-01MD-3* (3), *FsF-I2-01M-2* (1) and *FsF-R1.2-01M-2* (5).

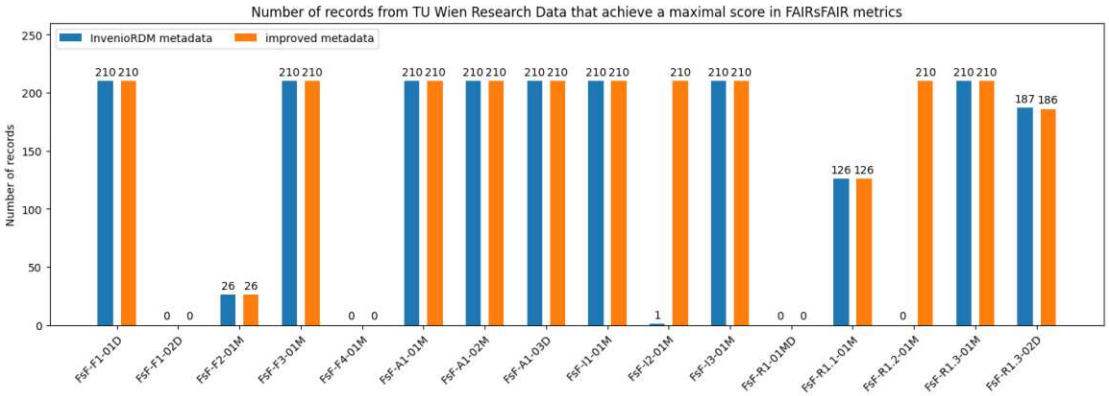

Figure 4.5: Number of records from TU Wien Research Data that achieve a maximal score in each of the FAIRsFAIR metrics. A juxtaposition of results between the metadata from InvenioRDM and the newly created format of metadata.

Nevertheless, even though we did not improve the test results of *FsF-R1-01MD-2* and *FsF-R1-01MD-3* for all records and we have not attempted to introduce any changes related to the *FsF-R1-01MD-4* test, the FUJI results with regard to FAIRsFAIR metrics are higher for our newly proposed format of the metadata. As depicted in Figure 4.5, the score has been improved for metrics: *FsF-I2-01M* and *FsF-R1.2-01M*. Furthermore, even though no record got a perfect score in the metric *FsF-R1-01MD*, more records achieve 3 out of 4 points than when using metadata from InvenioRDM. This is due to the improvements related to the *distribution* attribute. The mean score in this metric is 2.83 for the new standard of the metadata, compared to 1.73 for the metadata extracted from InvenioRDM.

4.4 Discussion

The improvements with regard to singular FAIRsFAIR metrics and FUJI tests contribute directly to higher FAIRness scores. In total, every record from the test instance of TU Wien Research Data repository achieved a higher score with the proposed metadata format compared to metadata extracted directly from InvenioRDM. The biggest observed change was 16.66 percentage points, followed by 12.5 and 8.33. As depicted in Figure 4.6, the highest achieved score was 87.5%, compared to 75% when using metadata from InvenioRDM.

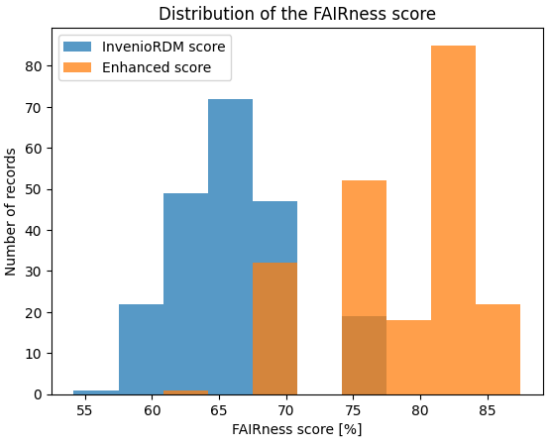

Figure 4.6: The distribution of the FAIRness score (in %) of records in the TU Wien Research Data measured by the FUJI tool. A juxtaposition of results between the metadata from InvenioRDM and the newly created format of metadata.

However, the observed increase in the FAIRness score is strictly related to the meticulous choice of LOD vocabularies and annotations. The application of the proposed LOD framework on its own did not influence the evaluation results returned by FUJI. It is due to the experiments described in Section 4.3 that we knew how to structure the metadata in a way that would increase its FAIRness, not because we followed the FDO specification in our implementation. This is a natural consequence of how we decided to measure the effectiveness of our solution—we used a pre-existing tool to achieve a quantifiable result. FUJI, however, is not equipped with mechanisms that would let it recognize the parts of such an infrastructure that we created.

Nonetheless, the usage of the proposed framework is a theoretical FAIRness improvement. This FDO framework is build around the idea of machine-actionability and interoperability. It is, thus, designed to be understandable by machines. Moreover, the framework is a concrete implementation based on the FDO specification and might be used as an example of how a FAIR system should be designed for any future projects in this area.

Additionally, the work presented in this section can be interpreted as a recommendation for the developing communities of both the InvenioRDM and FUJI projects on how to better represent the idea of FAIRness. From one side, FUJI can be used as a FAIRness oracle and the information about the *distribution* and *wasDerivedFrom* attributes might inspire the new versions of the InvenioRDM project, where the mentioned properties would be included in records' metadata by default. From the other side, if one assumes a position that supports the way in which InvenioRDM exports its records' metadata, it might also be a recommendation for FUJI that the lack of the aforementioned attributes

does not necessarily have to result in a decrease of the FAIRness score. Furthermore, we believe that the way FUJI interprets mime types and sizes of scanned files should be further investigated.

CHAPTER 5

Conclusion

In this thesis, we presented a concrete implementation of a FAIR Digital Objects system following the FDO specification. Our solution leverages the Cordra open source Digital Object management software as a backend for the FDOs. We chose to use Cordra due to it being lightweight and highly customizable (see Section 2.5). We introduced a Cordra-like schema object shared between every FDO in our project and 6 core entities that together make up for an FDO framework that can be fine-tuned for multiple purposes. This has been presented in Section 3.2. In Section 3.3, we also proposed a carefully chosen set of FDOs that make it possible to represent the metadata of records from a research data repository in our framework—the *Research Metadata Extension*.

Additionally, we conceptualized how the metadata from a given research data repository should be migrated to an FDO backend, where it is modelled using appropriate FDOs. We presented this procedure in Section 3.4.

Secondly, we introduced a tool—the *Migration Assistant*—that performs the migration procedure between a given instance of InvenioRDM and allows for a straightforward evaluation of the newly created format of the metadata using the chosen assessment tool—FUJI. After migrating the metadata of records from the TU Wien Research Data, we applied the automated tool FUJI to measure FAIRness of our proposed metadata representation. The achieved scores are higher than the ones acquired for metadata exported directly from TUWRD. However, this increase of the metadata FAIRness is strictly related to a careful choice of attribute names from well-known Linked Open Data vocabularies. The outcomes of the evaluation process we have carried out for the purpose of this thesis has been presented in Chapter 4.

We believe that the set of properties that were used to increase the FAIRness score of the metadata together with other outcomes of this thesis may spark a discussion in the development communities of both the InvenioRDM and FUJI projects on how to better

represent the idea of FAIRness. Furthermore, the proposed FDO framework might be an inspiration to any other FDO system implementations that may arise in the future.

The following sections aim to reiterate the research questions mentioned in Section 1.1 and also reference exactly where and how they have been answered. Additionally, we provide our outlook on some future work that might be related to this thesis.

5.1 Research questions

Please find a detailed revision of the research questions mentioned in Section 1.1 below.

5.1.1 What are the appropriate FDO profiles and attributes that can be used to represent records in a research data repository as FDOs?

As an important part of this thesis, we have designed and developed a novel FAIR Digital Objects framework following the FDO specification. Our direct contributions are:

- a set of required attributes in every FDO profile and attribute in our framework in form of a Cordra Schema (Section 3.2.1);
- six core FDOs that together with the Schema make up for a novel FDO framework that can be fine-tuned for multiple purposes and integrated with a LOD vocabulary of choice (Section 3.2.1);
- a set of additional FDOs that extend the core entities of our framework, so that it would be possible to model the metadata of InvenioRDM records as FDOs—the *Research Metadata Extension* (Section 3.3);
- a chosen set of LOD annotations that increase the machine-actionability of our system even further (Section 3.3.1).

Thus, a detailed overview of the appropriate FDO profiles and attributes that can be used to represent records in a research data repository as FDOs can be found in Chapter 3.

5.1.2 What is the conceptual architecture of a research data repository based on InvenioRDM that represents datasets as FDOs?

Another critical aspect of this thesis was to design and implement a solution that would serve as a layer between an InvenioRDM-based repository and the Cordra digital object management software. We have developed the application called *Migration Assistant* that allows for a seamless migration of metadata records from an instance of InvenioRDM to Cordra and an ability to test the newly created metadata using the FUJI automated tool. Together with the InvenioRDM and Cordra instances, this application is an example of

how a research data repository based on InvenioRDM that represents datasets as FDOs should be structured.

The conceptual architecture of a research data repository based on InvenioRDM that represents datasets as FDOs has been presented in Section 4.1.

5.1.3 What are the advantages of adopting the proposed solution with regard to the FAIR principles?

The solution developed for the purposes of this thesis should be evaluated in terms of satisfying the FDO specification and the impact it has on the FAIRness of the records in an existing repository.

First of all, our FDO framework has been developed strictly following the FDO specification, therefore extending an existing repository with the provided solution is in itself an advantage, because the repository records will become machine-actionable in a way that follows the newest industry standard—the FAIR Digital Objects. This elevates the overall FAIRness status of the data in the extended repository.

Second of all, the solution does not negatively impact the FAIRness level of the records in target data repositories. Having tested 210 records from the test instance of TU Wien Research Data using FUJI, we established that the FAIRness score increases after the application of the proposed solution for every tested record. We provide a detailed report regarding assessment in the chapter 4.

5.1.4 Limitations and future work

The solution created for the purpose of this thesis is a means to increase the FAIRness of records in a research data repository. We believe that future work that aims to achieve similar goals might benefit from the research that we carried out. First of all, our solution could be seamlessly integrated with the InvenioRDM project as an additional extending module, for instance. Secondly, similar tools might be developed for other research repository frameworks based on what we presented in this thesis.

However, even though we have proved that the usage of our solution results in an increase in the FAIRness of InvenioRDM records' metadata, we would like to point out again that it is strictly related to our meticulous choice of the LOD annotations that are accepted by the chosen assessment tool—FUJI. Similar results might be achieved without the usage of our novel FDO framework.

Furthermore, our findings might serve as a recommendation for the developing communities of both the InvenioRDM and FUJI projects on how to better represent the idea of FAIRness. We sum up these recommendations in the following points:

1. From one side, if one treated FUJI as a FAIRness oracle, one might consider adding the *wasDerivedFrom* (or another annotation from PROV-O) property to the records'

metadata in InvenioRDM. This in turn, would increase the score in terms of the *FsF-I2-01M* and *FsF-R1.2-01M* metrics. Moreover, the addition of the *distribution* property to the metadata of all record types (not only to the metadata of records of type *dataset*) would increase the score when it comes to the *FsF-R1-01MD* metric. Important to note is that the size of the files listed in the *distribution* array should be provided in bytes.

2. From the other side, if one considered metadata exported from InvenioRDM as a perfect example of FAIR metadata, one may consider changing FUJI's functionality. Maybe the lack of PROV-O annotations and the *distribution* property should not always result in a decrease in the FAIRness score.

3. Regardless of the point of view, however, it should be revisited how FUJI parses file types and sizes. Our tests have shown that certain mime types are not recognized by FUJI and the tool does not handle the existence of empty files (Section 4.3).

APPENDIX A

Source code of the solution

Our implementation of the *Migration Assistant* tool has been published here: `https://doi.org/10.5281/zenodo.15309020`. The FDO-configuration schema, the core FDOs and the rest of the necessary FDOs are published there as well. The FDO-related files are in the *migration_assistant/migration/cordra_schemas* directory.

APPENDIX B

Test results

The results of our tests have been published here: `https://doi.org/10.5281/zenodo.15309053`. The repository contains the following assets:

- *invenio_scores*—a directory that contains FUJI test results for the metadata extracted from InvenioRDM;
- *enhanced_scores*—a directory that contains FUJI test results for the newly proposed metadata format;
- *perform_tests.ipynb*—a Jupyter Notebook file, where the tests have been carried out;
- *visualise_results.ipynb*—a Jupyter Notebook file, where the data is visualized (graphs, statistical properties);
- *FUJI_tests_report.ipynb*—a Jupyter Notebook file with the detailed reports about certain FUJI tests. This file contains information relevant to Section 4.

Overview of Generative AI Tools Used

The usage of generative AI tools[1] for the purpose of this thesis was to create a *bibtex* entry for [BL06] as the author was unsure how to create such entries when he started writing the thesis text. The same tool has been used to aid the researcher in preparing the abstract and translating it to german. Apart from this, no generative AI tools have been used to write any other part of this thesis.

[1] `https://chatgpt.com/` has been used.

List of Figures

List of Tables

List of Listings

Glossary

Bing search engine owned by Microsoft. 41

CERN European Organization for Nuclear Research. 20

Cordra an open source Digital Object management software. xi, xiii, 2–6, 9, 19, 20, 22, 23, 25–27, 31, 32, 34–40, 59, 61–63, 79, 80, 89, 93

DataCite a global community that attempts to ensure that research outputs and resources are openly available and connected. 14, 42, 64

DONA foundation non-profit organization from Geneva. 19, 25

Dryad Digital Repository a nonprofit membership organization that is committed to making data available for research and educational reuse now and into the future. 1, 16, 17

Dublin Core a general purpose metadata vocabulary for describing resources of any type. 14

European Union a political and economic union of 27 member countries that are located primarily in Europe. 1, 10, 11

FAIR the set of principles that state that data should be: findable (F), accessible (A), interoperable (I) and reusable (R). xi, xiii, 1–3, 5, 6, 9–11, 16, 18, 21, 22, 42, 61, 77, 81, 82

FAIRness the extent to which a certain entity is FAIR (satisfies the FAIR guidelines). xi, xiii, xv, 2–7, 9–11, 13, 16, 17, 21, 24, 40–43, 58, 61–65, 67, 69, 70, 77–82, 89, 90

FAIRsFAIR a project—Fostering Fair Data Practices in Europe—that aims to supply practical solutions for the use of the FAIR data principles throughout the research data life cycle. xi, xiii, 2, 3, 5, 6, 9, 11–13, 17, 19, 21, 26, 63, 64, 76, 77, 89–91

Flask a framework for developing Python web applications. 62

FUJI a service that offers a programmatic assessment of the FAIRness of research datasets. xi, xiii, 3, 5–7, 13, 16, 17, 21, 41, 61–71, 76–82, 89–91

GO FAIR a bottom-up, stakeholder-driven and self-governed initiative that aims to implement the FAIR data principles. 18

Google an American organization that owns one of the most popular search engines of the same name. 41

InvenioRDM a research data repository framework which is an open source collaboration between many research institutions, such as CERN, the Northwester University and TU Wien. xi, xiii, xv, 3–7, 9, 20–22, 42, 44, 53, 61–65, 67, 69, 70, 76, 77, 79–82, 89, 90

JSON-LD a method of encoding linked data by using the JSON format. 9, 14, 19, 26, 40–43, 58, 63, 69

Northwestern University a private research university in Evanston, Illinois, United States. 21

Open Science a global effort to make scientific research and its outcomes accessible to everyone. 1

Open Science Framework a community with a mission to increase openness, integrity, and reproducibility of research. 1, 16

PROV a W3C specification on how to model provenance. 15

PROV-O an owl ontology that models the concepts from PROV. 15, 40, 42, 43, 66, 70, 81, 82, 93

Python a popular multipurpose programming language. 62

RDFa (Resource Description Framework in Attributes) a W3C Recommendation that adds a set of attribute-level extensions to HTML, XHTML and XML-based document types for embedding rich metadata within web documents. 14

schema.org a broad vocabulary created in 2011 by Google, Yahoo, Yandex and Bing, which, lately, has become one of the most popular vocabularies worldwide. 18, 40–43, 69, 93

Semantic Web a web of machine-readable data. 9, 10

Signposting a technique for embedding machine-readable links in HTML that describe the relationships between different resources. 21

WorldFAIR a project that sets out to produce recommendations, interoperability frameworks and guidelines for FAIR data assessment. 5, 11

Yahoo an American company known for their search engine—Yahoo Search. 41

Yandex a Russian company that provides internet products, such as their search index—Yandex Search. 41

Zenodo open-access research data repository developed by CERN and OpenAIRE. 1, 16, 21, 22

Acronyms

CDIF Cross-Domain Interoperability Framework. 11

CNRI Corporation for National Research Initiatives. 19, 25

CODATA Committee on Data of the International Science Council. 1, 10, 11

DCAT Data Catalog Vocabulary. 41

DCMI Dublin Core Metadata Terms. 40, 42, 43, 69, 93

DOA Digital Object Architecture. 4, 19, 23, 25

DOIP Digital Object Interface Protocol. 4, 19, 25, 62

DOIRP Digital Object Identifier Resolution Protocol. 19, 26

FDO FAIR Digital Object. xi, xiii, xv, 2–6, 18, 19, 23–27, 29, 31–42, 44–46, 48–54, 56–58, 61, 63, 69, 77, 79–81, 83, 89, 93

FDOs FAIR Digital Objects. xi, xiii, 2–4, 6, 9, 10, 18, 22–25, 27–29, 31, 32, 34, 36–38, 40, 41, 43, 44, 53, 54, 56–59, 61–63, 79–81, 83, 89

IRI Internationalized Resource Identifier. 13

JRC Joint Research Center. 10

LOD Linked Open Data. 4, 5, 9–11, 18, 19, 26, 40–42, 44, 58, 77, 79–81, 93

OWL Web Ontology Language. 9, 16

PID Persistent Identifier. 18, 27, 64

RDA Research Data Alliance. 1, 10, 11, 15, 16, 21

RDF Resource Description Framework. 9, 14, 40

RO-Crate Research Object Crate. 11

SHACL Shapes Constraint Language. 16

SKOS Simple Knowledge Organization System. 16, 41

SPARQL SPARQL Protocol and RDF Query Language. 16

SPDX Software Package Data Exchange. 15

TUWRD TU Wien Research Data. 1, 3, 5, 6, 17, 21, 22, 61, 63–65, 70, 76, 77, 79, 81, 89, 90

URI Uniform Resource Identifier. 14

URL Uniform Resource Locator. 13, 66–68, 70, 71

UUID Universally Unique Identifier. 13

W3C World Wide Web Consortium. 9, 41

WWW World Wide Web. 9

Bibliography

[ABJ22] Emna Amdouni, Syphax Bouazzouni, and Clement Jonquet. O'faire makes you an offer: metadata-based automatic fairness assessment for ontologies and semantic resources. *International Journal of Metadata, Semantics and Ontologies*, 16(1):16–46, 2022.

[AHG20] Dean Allemang, Jim Hendler, and Fabien Gandon. *Semantic Web for the Working Ontologist: Effective Modeling for Linked Data, RDFS, and OWL*, volume 33. Association for Computing Machinery, New York, NY, USA, 3 edition, 2020.

[BHBL09] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data - the story so far. In *International Journal on Semantic Web and Information Systems (IJSWIS)*, 2009.

[BL06] Tim Berners-Lee. Linked data - design issues, 2006. Accessed: 2025-01-25.

[BLHL01] Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. *Scientific American*, 284(5):34–43, 2001.

[Car23] Matthew Carson. Inveniordm: Supporting fair principles and policies, June 2023.

[CBB+24] Simon Cox, David Browning, Alejandra Gonzalez Beltran, Peter Winstanley, Andrea Perego, and Riccardo Albertoni. Data catalog vocabulary (DCAT) - version 3. W3C recommendation, W3C, August 2024.

[CLW14] Richard Cyganiak, Markus Lanthaler, and David Wood. RDF 1.1 concepts and abstract syntax. W3C recommendation, W3C, February 2014.

[CML+25] Blanchi Christophe, Hellström Maggie, Lannom Larry, Pfeil Andreas, Schwardmann Ulrich, and Wittenburg Peter. Implementation of attributes, types, profiles and registries, January 2025.

[Cor23] Corporation for National Research Initiatives (CNRI). *Cordra® Software Technical Manual*, 2023. Accessed: 2025-03-22.

[CWJ+19] Daniel J. B. Clarke, Lily Wang, Alex Jones, Megan L. Wojciechowicz, Denis Torre, Kathleen M. Jagodnik, Sherry L. Jenkins, Peter McQuilton, Zachary Flamholz, Moshe C. Silverstein, Brian M. Schilder, Kimberly Robasky, Claris Castillo, Ray Idaszak, Stanley C. Ahalt, Jason Williams, Stephan Schurer, Daniel J. Cooper, Ricardo de Miranda Azevedo, Juergen A. Klenk, Melissa A. Haendel, Jared Nedzel, Paul Avillach, Mary E. Shimoyama, Rayna M. Harris, Meredith Gamble, Rudy Poten, Amanda L. Charbonneau, Jennie Larkin, C. Titus Brown, Vivien R. Bonazzi, Michel J. Dumontier, Susanna-Assunta Sansone, and Avi Ma'ayan. Fairshake: Toolkit to evaluate the fairness of research digital resources. *Cell Systems*, 9(5):417–421, 2025/04/28 2019.

[DH20] Anusuriya Devaraju and Robert Huber. F-uji - an automated fair data assessment tool, October 2020.

[DHM+22] Anusuriya Devaraju, Robert Huber, Mustapha Mokrane, Patricia Herterich, Linas Cepinskas, Jerry de Vries, Herve L'Hours, Joy Davidson, and Angus White. Fairsfair data object assessment metrics, April 2022.

[DNPR18] Fariz Darari, Werner Nutt, Giuseppe Pirrò, and Simon Razniewski. Completeness management for rdf data sources. *ACM Trans. Web*, 12(3), July 2018.

[ea22] Soiland-Reyes et al. Packaging research artefacts with ro-crate. *Data Science*, 5(2):97–138, 2022.

[Fay10] Ed Fay. Repository software comparison: Building digital library infrastructure at lse. *Ariadne*, (64), 2010.

[FMS+23] Janine Felden, Lars Möller, Uwe Schindler, Robert Huber, Stefanie Schumacher, Roland Koppe, Michael Diepenbroek, and Frank Oliver Glöckner. Pangaea - data publisher for earth & environmental science. *Scientific Data*, 10(1):347, 2023.

[GBB+24] Arofan Gregory, Darren Bell, Dan Brickley, Pier Luigi Buttigieg, Simon Cox, Michelle Edwards, Fils Doug, Luis Gerardo Gonzalez Morales, Pascal Heus, Simon Hodson, Chifundo Kanjala, Yann Le Franc, Lauren Maxwell, Laura Molloy, Steve Richard, Flavio Rizzolo, Peter Winstanley, Lesley Wyborn, and Adrian Burton. Worldfair (d2.3) cross-domain interoperability framework (cdif) (report synthesising recommendations for disciplines and cross- disciplinary research areas), May 2024.

[GBM16] R. V. Guha, Dan Brickley, and Steve Macbeth. Schema.org: evolution of structured data on the web. *Commun. ACM*, 59(2):44–51, January 2016.

[GCPV21] Daniel Garijo, Oscar Corcho, and Maria Poveda-Villalón. Foops!: An ontology pitfall scanner for the fair principles. 2980, 2021.

[Gil24] MATHIEU Gilles. A general presentation of the european open science cloud (eosc), November 2024.

[GPV20] Daniel Garijo and María Poveda-Villalón. Best practices for implementing fair vocabularies and ontologies on the web, 2020.

[GRDL+23] Alban Gaignard, Thomas Rosnet, Frédéric De Lamotte, Vincent Lefort, and Marie-Dominique Devignes. Fair-checker: supporting digital resource findability and reuse with knowledge graphs and semantic web standards. *Journal of Biomedical Semantics*, 14(1):7, 2023.

[Gro20] FAIR Data Maturity Model Working Group. Fair data maturity model. specification and guidelines, June 2020.

[HJC+18] Simon Hodson, Sarah Jones, Sandra Collins, Françoise Genova, Natalie Harrower, Leif Laaksonen, Daniel Mietchen, Rūta Petrauskaité, and Peter Wittenburg. Turning fair data into reality: interim report from the european commission expert group on fair data, June 2018.

[HTL+25] Lowenthal H, Austin T, Bonino Da Silva Santos LO, Chiarelli C, Cusinato A, Ferigato C, Friis-Christensen A, Kemper T, Perrotta D, and Wittwehr C. Jrc fair data guidelines. (KJ-01-25-098-EN-N (online)), 2025.

[IASK25] Andrew Iliadis, Amelia Acker, Wesley Stevens, and Sezgi Başak Kavakli. One schema to rule them all: How schema.org models the world of search. *Journal of the Association for Information Science and Technology*, 76(2):460–523, 2025.

[ICD+23] Anders Ivonne, Blanchi Christophe, Broder Daan, Hellström Maggie, Islam Sharif, Jejkal Thomas, Lannom Larry, Peters-von Gehlen Karsten, Quick Robert, Schlemmer Alexander, Schwardmann Ulrich, Soiland-Reyes Stian, Strawn George, van Uytvanck Dieter, Weiland Claus, Wittenburg Peter, and Zwölf Carlo. Fair digital object technical overview, April 2023.

[KBL+18] Robert E. Kahn, Christophe Blanchi, Laurence Lannom, Patrice A. Lyons, Giridhar Manepalli, Robert Tupelo-Schneck, and Sam Sun. Digital object interface protocol specification version 2.0. Technical report, DONA Foundation, November 2018.

[KBTS+22] Robert E. Kahn, Christophe Blanchi, Robert Tupelo-Schneck, Laurence Lannom, Patrice A. Lyons, Giridhar Manepalli, and Sam Sun. Digital object identifier resolution protocol specification version 3.0. Technical report, DONA Foundation, June 2022.

[KW06] Robert Kahn and Robert Wilensky. A framework for distributed digital object services. *International Journal on Digital Libraries*, 6(2):115–123, 2006.

[LRA+24] Matthias Liffers, Wendy Robertson, Jan Ashton, Isabel Bernal, Anusuriya Devaraju, Kirsten Elger, Vanessa Gabriel, Ted Habermann, Andrea Medina-Smith, Joseph Padfield, Jessica Parland von Essen, Anne Raugh, Mike Shallcross, Nicola Tarocco, Hana Vyčítalová, Alexander Whelan, Kelly Stathis, and Sara El-Gebali. Datacite metadata schema for the publication and citation of research data and other research outputs. version 4.6, 2024. Accessed: 2025-04-30.

[MB09] Alistair Miles and Sean Bechhofer. SKOS simple knowledge organization system reference. W3C recommendation, W3C, August 2009.

[Mol22] Laura Molloy. Worldfair project - announcement of project launch, June 2022.

[MTDF17] Pasquale Minervini, Volker Tresp, Claudia D'amato, and Nicola Fanizzi. Adaptive knowledge propagation in web ontologies. *ACM Trans. Web*, 12(1), August 2017.

[PSMP12] Peter Patel-Schneider, Boris Motik, and Bijan Parsia. OWL 2 web ontology language structural specification and functional-style syntax (second edition). W3C recommendation, W3C, December 2012.

[PVEAGC20] María Poveda-Villalón, Paola Espinoza-Arias, Daniel Garijo, and Oscar Corcho. Coming to terms with fair ontologies. In *Knowledge Engineering and Knowledge Management: 22nd International Conference, EKAW 2020, Bolzano, Italy, September 16–20, 2020, Proceedings*, page 255–270, Berlin, Heidelberg, 2020. Springer-Verlag.

[QBC13] Silvia Quarteroni, Marco Brambilla, and Stefano Ceri. A bottom-up, knowledge-aware approach to integrating and querying web data services. *ACM Trans. Web*, 7(4), November 2013.

[QT25] Vivian Yifei Qiu and Alex Wing Cheung Tse. Research trends and emerging themes of data repository for research data management services in higher education: a bibliometric analysis. In *Proceedings of the 2024 16th International Conference on Education Technology and Computers*, ICETC '24, page 295–303, New York, NY, USA, 2025. Association for Computing Machinery.

[SKL14] Manu Sporny, Gregg Kellogg, and Markus Lanthaler. JSON-ld 1.0. W3C recommendation, W3C, January 2014.

[TS22] Robert Tupelo-Schneck. An introduction to cordra., 2022.

[TTHS19] Steffen Thoma, Andreas Thalhammer, Andreas Harth, and Rudi Studer. Fuse: Entity-centric data fusion on linked data. *ACM Trans. Web*, 13(2), February 2019.

[Vig24] Guillaume Viger. Signposting in inveniordm, June 2024.

[WDA+16] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018, 2016.