Advancing FAIR practices in biomedical research

Challenges, methods and innovations for sustainable data management

César Bernabé

Advancing FAIR practices in biomedical research: challenges, methods and innovations for sustainable data management

César Henrique Bernabé

Typeset in LaTeX by César Bernabé
Cover design by César Bernabé
(Figure is AI-generated)

Printed by Ridderprint

Advancing FAIR practices in biomedical research: challenges, methods and innovations for sustainable data management

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op woensdag 26 november 2025
klokke 11.30 uur

door César Bernabé

geboren te Colatina, Brazilië

in 1992

Promotor:	Prof. dr. B. Mons	
Copromotor:	Dr. M. Roos	
Promotiecomissie:	Prof.dr. M.R. Spruit	
	Prof.dr. R. Cornet	(Amsterdam UMC)
	Prof.dr. M.E.H. van Reisen	
	Dr. R.S.S. Guizzardi	(Universiteit van Twente)
	Prof.dr. V.E.S. Souza	(Universidade Federal do Espírito Santo)

Dedico este trabalho aos meus avós, Mário e Zulmira.
Eu sei que vocês zelam por mim aí do céu.

“Knowledge emerges only through invention and re-invention, through the restless, impatient, continuing, hopeful inquiry human beings pursue in the world, with the world, and with each other.”

— Paulo Freire.

Contents

Chapter 1

Introduction

What creates harmonisation? In an orchestra, the harmonious collaboration of diverse musical instruments creates a cohesive symphony. This synchronisation is possible because all musicians agree to follow a common musical score under the guidance of a conductor. Similarly, for research data to be valuable for integration and reuse, it must harmonise with other data. This requires adherence to conducting principles of good data management. Without principles that ensure that knowledge resources are built harmoniously, the result would be a combination of incompatible data notes.

In data management, the FAIR guiding principles [1] are recommended for harmonising data and other resources. However, their implementation can be complex due to the principles' multifaceted nature and the variety of ways in which they can be realised. This thesis aims to address these complexities by first identifying the main challenges of the process of making resources FAIR (FAIRification). It then leverages existing approaches from related fields to tackle different aspects of the FAIRification process [2]. Specifically, the thesis describes FAIR training formats tailored to different types of stakeholders. Subsequently, it introduces a goal-based method to support FAIRification planning and the identification of FAIRification objectives. Finally, it examines ontological artefacts to understand their impact on the quality of FAIR (meta)data models and the resulting FAIRified resources.

1.1 Data as instruments to foster research

Some data-driven research projects generate, manage, and publish new data, while others reuse existing knowledge for further analysis, often creating additional data in the process. In cases where data are reused, it is common to integrate various sources from different studies to create a more comprehensive and meaningful knowledge base. However, reusing research data can be challenging due to variations in languages, data collection and storage methods, formats (e.g. CSV, RDF), and interpretations of data concepts (e.g. treatment as a guideline vs. treatment as a process). These differences complicate data reuse, especially for researchers who perform machine-assisted (semi-)automatic integration of large data. A report from the European Commission estimated (in 2018) that the lack of reusable data results in an annual economic cost of at least 10.2 billion euros [3].

Consequently, to address the issue of non-reusable data, several funding [4] and government bodies [5] have been raising awareness of the importance of the FAIR guiding principles to make resources Findable, Accessible, Interoperable and Reusable. Most funders now require researchers to document their methodology for making resources FAIR as part of a Data Management Plan [4].

1.2 FAIR principles for harmonious data management

The FAIR principles were published in 2016 by a group of stakeholders representing academia, industry, funding agencies, and scholarly publishers. These guiding principles state that research resources, such as data and associated metadata, should be made FAIR for both humans and machines [1]. The principles are depicted in Table 1.1, and further discussed next.

Findability Unlike the clear arrangement of musicians in an orchestra, where each section (e.g., violins) has a defined, visible position, research resources are often dispersed across multiple, poorly indexed locations. This lack of organisation makes it difficult for researchers to locate the resources they need. Therefore, it is essential to make research resources automatically findable by machines. Findability is composed of four sub-principles: using globally unique and persistent identifiers for resources and their metadata (principle F1); describing resources with rich metadata (F2); explicitly

Table 1.1: Description of the FAIR principles, from [1].

ID	Description
F1	(Meta) data are assigned globally unique and persistent identifiers
F2	Data are described with rich metadata
F3	Metadata clearly and explicitly include the identifier of the data they describe
F4	(Meta)data are registered or indexed in a searchable resource
A1	(Meta)data are retrievable by their identifier using a standardised communication protocol
A1.1	The protocol is open, free and universally implementable
A1.2	The protocol allows for an authentication and authorisation procedure where necessary
A2	Metadata should be accessible even when the data is no longer available
I1	(Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
I2	(Meta)data use vocabularies that follow the FAIR principles
I3	(Meta)data include qualified references to other (meta)data
R1	(Meta)data are richly described with a plurality of accurate and relevant attributes
R1.1	(Meta)data are released with a clear and accessible data usage license
R1.2	(Meta)data are associated with detailed provenance
R.13	(Meta)data meet domain-relevant community standards

including the resource's identifier in the metadata (F3); and ensuring that resources are indexed in searchable platforms (F4).

Accessibility When arranging musicians in an orchestra, it is essential to clearly identify each musician's specialisation to ensure, for instance, that violin players are assigned to violins. Similarly, a cello player would not be assigned to play the flute, as they likely lack the necessary skills (credentials) to access and play that instrument. In the context of FAIR resources, accessibility requires clear access conditions to ensure that only authorised agents (i.e., humans and/or machines) can access resources under specific conditions. The accessibility principle recommends the use of open, free, and universally implementable standardised communication protocols (A1, A1.1) and the provision of clear authentication and authorisation instructions where necessary (A1.2). Accessibility is also enhanced by ensuring the longevity of metadata (A2).

Interoperability When organising a concert, violin players can, for instance, use instruments from different manufacturers, as long as all instruments are tuned to produce the same notes. In a data context, interoperability is similarly achieved, as it does not require that all 'violins are identical' (i.e. only one standard is followed), nor that 'all instruments are violins' (i.e. different standards fit to different contexts), but rather ensuring that 'different instruments are compatible with the overall harmony' (i.e. standards are compatible among themselves). For FAIR resources, interoperability is enhanced by publishing resources and their metadata in common knowledge representation formats (I1), using vocabularies that also follow the FAIR principles (I2), and including qualified references to other resources (I3).

Reusability When writing a piece, the composer must include detailed information to allow others to perform it. For instance, the composer must provide notes for different types of instruments and specify the tempo to be followed. Similarly, reusability involves providing rich information so as to enable both humans and machines to make well-informed decisions about effectively reusing FAIR resources. The FAIR principles address reusability by recommending that resources be described with relevant attributes such as clear licencing (R1.1) and provenance information (R1.2), while following domain-relevant community standards (R1.3).

1.3 FAIRification: the orchestration of the FAIR principles

FAIRification is the process of making resources FAIR [2]. It can be compared to the process of composing a symphony. When writing a piece, the composer must follow several steps to produce her/his piece, some of these steps are more creative and others more technical. For example, the composer must arrange the notes to create harmony and then transcribe these notes into different scores for various musicians and instruments. These scores should be written in a standard format so that others can follow them. Similarly, in the FAIRification process, the resource to be made FAIR is analysed, modelled according to different standards, and made available for diverse reuse scenarios. The FAIRification process is organised by FAIRification workflows [2]. While some workflows are tailored to specific communities (e.g. workflows for health research [6]), others are designed to be generic for broader applicability. For example, the generic FAIRification workflow [2] defines seven steps, organised into three phases, as shown in Figure 1.1.

In the pre-FAIRification phase, the FAIRification objectives are identified (step 1), and the data and metadata are analysed (steps 2 and 3). Examples of (meta)data analysis include investigating representation formats and the meaning (semantics) of (meta)data elements. In the FAIRification phase, the semantic models of the data and the metadata are defined (steps 4a and 4b). These include reusing existing models or generating a semantic model through ontology-driven conceptual modelling. Subsequently, the semantic models are used to make the data and metadata linkable (steps 5a and 5b). The linked (meta)data is then hosted using community relevant formats to make it available to humans and machines. In the post-FAIRification phase, the FAIR resource is assessed to ensure that it meets the FAIRification objectives. The generic

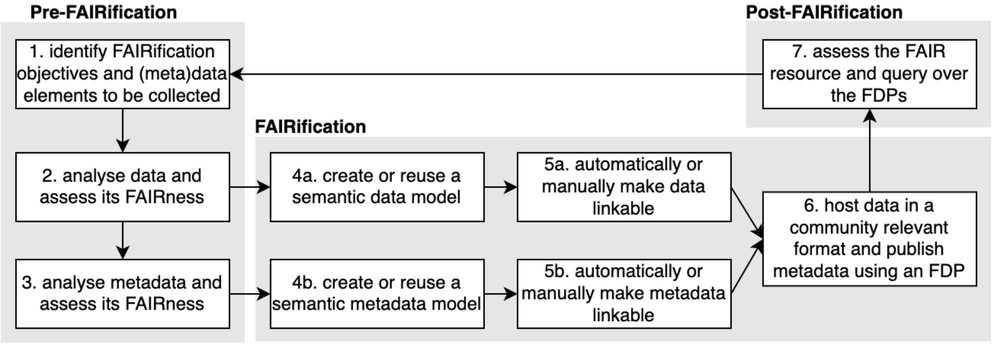

Pre-FAIRification
1. identify FAIRification objectives and (meta)data elements to be collected
2. analyse data and assess its FAIRness
3. analyse metadata and assess its FAIRness
FAIRification
4a. create or reuse a semantic data model
4b. create or reuse a semantic metadata model
5a. automatically or manually make data linkable
5b. automatically or manually make metadata linkable
6. host data in a community relevant format and publish metadata using an FDP
Post-FAIRification
7. assess the FAIR resource and query over the FDPs

Figure 1.1: The generic FAIRification workflow [2].

FAIRification is designed for retrospective FAIRification, when existing resources are made FAIR. Other workflows, such as the de novo FAIRification workflow [6], are designed for prospective FAIRification, when resources are made FAIR upon creation.

1.4 FAIR and rare diseases research

The FAIR principles have gained traction in many fields, and are particularly relevant to rare disease research. Although rare diseases may seem less prevalent when considered in isolation, they are highly relevant on a global scale. For example, while a single hospital may collect data from only a few rare disease patients, the global prevalence of rare diseases has been estimated to be between 280 and 473 million people in 2023, based on the existence of approximately 3,585 rare diseases [7]. Consequently, data collected locally in hospitals and specialised centres is insufficient for data-driven research. To address this, it is necessary to merge data from different sources, often from different institutions and countries. However, the integration of multiple resources faces several barriers, both technical and legal. On the technical side, resources need to be findable, standards need to be harmonised and metadata need to be clearly defined for humans and machines. Legally, it is important to ensure that sensitive data is only accessed by authorised agents, and that data and other resources are used under authorised conditions. Therefore, the FAIR principles play an essential role in supporting rare disease research by facilitating the integration and reuse of data, and clarifying the conditions for access and reuse.

Currently, several initiatives rely on the FAIR principles to promote the harmonisation of rare disease data (e.g. [8]). For example, the European Joint Programme on

Rare Diseases (EJP RD) [8], a five-year initiative conducted from 2020 to 2024, involved over 130 institutions from EU member states and associated countries to create an ecosystem of FAIR resources. The EJP RD aimed at providing a technical infrastructure to discover, query and eventually access data from different types of sources such as patient registries, biobanks, genomics and multi-omics repositories, knowledge bases and other related resources (such as animal models and cell line libraries) in a coordinated manner.

One of the efforts created by the EJP RD to support the FAIRification of rare disease resources was the FAIRification Stewards group, a team established to support the EJP RD partners in the FAIRification process of their resources. The objectives of the Stewards included identifying the needs of the participating institutions, providing advice and promoting convergence in FAIR implementation decisions. The author of this thesis has been part of the FAIRification Stewards team since its creation, and the experience gained while working with this team has motivated the research presented herewith.

During their activities to support the FAIRification of different rare disease resources, the FAIR Data Stewards team catalogued the challenges faced during the process, and proposed specific solutions to those. While solutions were proposed to most challenges (i.e. in the context of EJP RD), there was still the need to address them in a scientific manner. Among the catalogued challenges (cf. Section 2.4), it was observed that: (i) a shortage of FAIR experts and the need for continuous training persist, (ii) difficulties arise in defining FAIRification goals prior to implementation, (iii) challenges exist in identifying and adopting community-relevant standards and appropriate semantic models, and (iv) addressing legal concerns, such as privacy and ethics, remains a complex aspect of developing a FAIR solution. Hence, the first three observations are discussed in this thesis, and solutions to them are investigated and designed. The fourth one is out of the scope of this thesis given its need for specific expertise (e.g. legal advisors).

1.5 Research gaps in FAIR(ification)

The challenges identified during the EJP RD, arising from its international and diverse context, are broadly relevant across various domains. The lack of available expertise poses a significant problem for the adoption of FAIR principles, as it slows down the process and reduces its effectiveness. In most cases, experts in this field need to be familiar with the domain-specific aspects, as well as the theoretical and technical aspects required to realise the principles, which makes finding qualified individuals challenging [9]. Additionally, the scarcity of technical experts can result in the inefficient use of resources and funding, ultimately preventing organisations from fully leveraging the benefits of FAIR data.

FAIRification encompasses identifying FAIRification objectives, which is also an important aspect that impacts the entire FAIRification process and the FAIRness (a degree to which a resource adheres to the FAIR principles) of the resulting resource. This is because different implementation choices can be made considering the FAIRification objectives. Consequently, this step must be executed thoroughly to effectively guide the subsequent tasks. However, identifying FAIRification objectives can be complex, as it requires assessing the current state of the resource to be made FAIR and their surrounding environment, such as the organisational infrastructure and policies. Additionally, it involves identifying standards, indexing repositories, access conditions, file formats and licensing that align with the needs of the community or stakeholders who will reuse the FAIRified resource. This can be particularly overwhelming, especially for teams with limited experience in FAIRification.

Similarly, designing (meta)data semantic models is a critical step because it essentially defines the final shape of the FAIRified resource. This impacts, for instance, how widely the resource will be reused, as poor semantic models hinder understandability and consequently reusability. Therefore, semantic models should be unambiguous, clear, and adaptable to other contexts [10]. Additionally, for machines to be able to automatically interpret the data, the semantic model must be annotated with reliable ontologies and be richly axiomatised. These challenges are compounded by the intrinsic complexity of modelling various domains, especially when the domain itself is complex (e.g. modelling genetic information). For instance, determining which elements to include in the semantic model, the level of refinement for these elements, and the types of relationships allowed among elements can be a difficult task. If the metadata and data models are not well understood, clearly defined, and flexible enough to be managed by others, the FAIR resource will not fulfil its ultimate purpose of reusability [10].

1.6 Leveraging on other research genres to support FAIRification

While some FAIRification challenges might be novel to research on FAIR, similar issues have already been investigated by other communities. The requirements engineering, a subfield of software engineering research, studies methods to properly identify objectives and requirements that guide software development [11, 12]. Similarly, prior research on ontologies builds on theories from cognitive science, philosophy, logics and linguistics to develop artefacts and methods to guide ontology engineering and conceptual model design [13, 14, 15].

Goal-modelling [12] is one of the state-of-the-art methods in requirements engineering to support the identification of objectives and requirements for software development. Goal-modelling methods include frameworks and modelling languages to support the identification, documentation and communication of goals among all interested stakeholders. The *iStar* framework [16], for instance, supports analysing the goals, motivations, and dependencies of actors within a system. *iStar* explores the rationale of stakeholders, detailing their goals, tasks, resources, and the ways they achieve their objectives.

Foundational ontologies [13, 14] are artefacts designed to provide the basic structure and concepts necessary for developing more specific domain ontologies and conceptual models. They serve as a high-level starting point for organising knowledge, ensuring consistency and interoperability across different systems and domains. For instance, OntoUML [17] is a conceptual modelling language designed for creating ontologically well-founded models, based on the underlying theories of the Unified Foundational Ontology (UFO) [18].

1.7 Aims and outline of this thesis

The primary aim of this research is twofold. First, to gain insight into FAIRification from the perspective of its target audience (i.e. researchers reusing FAIR data), identifying the most challenging aspects of the process. Consequently, the second aim of this thesis is to propose solutions to these challenges, drawing on results from other fields whenever possible, as building on already mature research contributes to the robustness of the results presented here.

In order to achieve its stated aim, the thesis is divided into four parts. These parts are illustrated in Figure 1.2 together with the relationships between them. Part

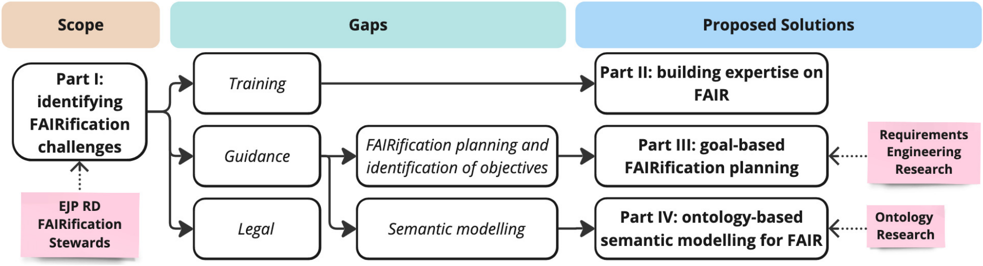

Scope
Gaps
Proposed Solutions
Part I: identifying FAIRification challenges
EJP RD FAIRification Stewards
Training
Guidance
Legal
FAIRification planning and identification of objectives
Semantic modelling
Part II: building expertise on FAIR
Part III: goal-based FAIRification planning
Part IV: ontology-based semantic modelling for FAIR
Requirements Engineering Research
Ontology Research

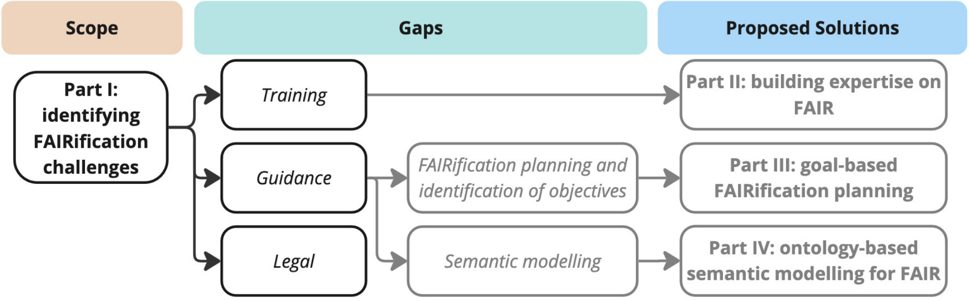
Figure 1.2: Outline of this thesis. The figure explains how gaps were identified from Part I, and how proposed solutions relate to these gaps. The pink stickers depict projects or research areas that informed a certain part.

I describes the EJP RD FAIRification Stewards team and their efforts to identify and catalogue FAIRification challenges in FAIRifying rare disease data. The main gaps identified in Part I relate to the need for training, guidance on FAIRification planning and semantic modelling, and the need to address legal constraints. Part II presents formats of training workshops tailored to different types of stakeholders (i.e. researchers, managers, and developers), thus addressing the training need. Part III describes a method to support FAIRification planning using goal modelling techniques, thus targeting the need for guidance on planning and identifying FAIRification goals. Part IV examines the use of foundational ontologies in the biomedical domain and in a machine learning (ML) related study, thus providing initial guidance for the design of semantic models for FAIRification. Finally, Chapter 7 summarises the main findings and provides an overall discussion of the work presented in this thesis.[1]

[1] This thesis has benefited from the use of AI-assisted tools for proofreading. All intellectual contributions, arguments, and interpretations remain those of the author of this thesis and his co-authors.

Part I

Identifying FAIRification challenges


Scope
Gaps
Proposed Solutions
Part I: identifying FAIRification challenges
Training
Guidance
Legal
FAIRification planning and identification of objectives
Semantic modelling
Part II: building expertise on FAIR
Part III: goal-based FAIRification planning
Part IV: ontology-based semantic modelling for FAIR

Topics covered in Part I of this thesis.

This part describes an effort to FAIRify sensitive and fragmented rare disease patient data, and the challenges identified in this process. Chapter 2 describes the challenges of making European Registry Networks (ERNs) FAIR within the context of EJP RD. During this initiative, 24 ERNs were supported by the FAIRification Stewards group throughout the FAIRification process. While Chapter 2 represents the primary source of experience and expertise that shaped the PhD research presented in this thesis, additional projects and initiatives also contributed (e.g. [19, 20, 21, 22]). Collectively, these experiences helped identify overarching challenges that are widely relevant to current FAIRification efforts. These challenges can be summarised as: (i) the need for greater specialised expertise and awareness (i.e. training), addressed in Part II; (ii) the need for clearer guidance on defining FAIRification objectives, tackled in Part III; and (iii) the semantic modelling of (meta)data, explored in Part IV.

Chapter 2

Towards FAIRification of sensitive and fragmented rare disease patient data

Challenges and solutions in European reference network registries

Bruna dos Santos Vieira*, César H. Bernabé*, Shuxin Zhang*, Haitham Abaza, Nirupama Benis, Alberto Cámara, Ronald Cornet, Clémence M. A. Le Cornec, Peter A. C. ’t Hoen, Franz Schaefer, K. Joeri van der Velde, Morris A. Swertz, Mark D. Wilkinson, Annika Jacobsen and Marco Roos

**BSV, CHB, and SZ share first authorship, having contributed equally to the research, as well as the drafting, review, and editing of the manuscript. NB, AC, CMALC, and JV also contributed to conducting the work described. All authors contributed by revising the manuscript and providing critical feedback.*

Orphanet Journal of Rare Diseases, 17, 436 (2022). https://doi.org/10.1186/s13023-022-02558-5

Abstract

Introduction Rare disease patient data are typically sensitive, present in multiple registries controlled by different custodians, and non-interoperable. Making these data Findable, Accessible, Interoperable, and Reusable (FAIR) for humans and machines at source enables federated discovery and analysis across data custodians. This facilitates accurate diagnosis, optimal clinical management, and personalised treatments. In Europe, twenty-four European Reference Networks (ERNs) work on rare disease registries in different clinical domains. The process and the implementation choices for making data FAIR ('FAIRification') differ among ERN registries. For example, registries use different software systems and are subject to different legal regulations. To support the ERNs in making informed decisions and to harmonise FAIRification, the FAIRification steward team was established to work as liaisons between ERNs and researchers from the European Joint Programme on Rare Diseases.

Results The FAIRification steward team inventoried the FAIRification challenges of the ERN registries and proposed solutions collectively with involved stakeholders to address them. Ninety-eight FAIRification challenges from 24 ERNs' registries were collected and categorised into "training" (31), "community" (9), "modelling" (12), "implementation" (26), and "legal" (20). After curating and aggregating highly similar challenges, 41 unique FAIRification challenges remained. The two categories with the most challenges were "training" (15) and "implementation" (9), followed by "community" (7), and then "modelling" (5) and "legal" (5). To address all challenges, eleven types of solutions were proposed. Among them, the provision of guidelines and the organisation of training activities resolved the "training" challenges, which ranged from less-technical "coffee-rounds" to technical workshops, from informal FAIR Games to formal hackathons. Obtaining implementation support from technical experts was the solution type for tackling the "implementation" challenges.

Conclusion This work shows that a dedicated team of FAIR data stewards is an asset for harmonising the various processes of making data FAIR in a large organisation with multiple stakeholders. Additionally, multi-levelled training activities are required to accommodate the diverse needs of the ERNs. Finally, the lessons learned from the experience of the FAIRification steward team described in this paper may help to increase FAIR awareness and provide insights into FAIRification challenges and solutions of rare disease registries.

2.1 Introduction

Rare diseases (RDs) are defined as life-threatening or chronically debilitating conditions that affect a low percentage of the population. In Europe, diseases are considered "rare" when their prevalence is less than 5 per 10,000 people [23]. Their low prevalence means that RD patient data is scarce and fragmented. Consequently, it is difficult to access sufficient data to support, for instance, research, drug development and improvements in outpatient care. The Orphanet, the National Organisation for Rare Diseases (NORD)[24], and other initiatives around the world have deemed it important to improve collaboration for research[25] and Open Science for RD [26]. Such initiatives make it easier for people with RDs to share their data. In fact, the importance of data sharing is consistently emphasised by RD patients themselves[27]. To help with research on RDs, the European Joint Programme on Rare Diseases (EJP RD) was set up in 2018 [8]. The programme aims to solve the problem of fragmented information and to build a research ecosystem that makes the best use of data and resources, thus benefiting people with RDs. The EJP RD project collaborates directly with the 24 European Reference Networks (ERNs) [28], which involve more than 900 highly specialised healthcare units from more than 130 institutions in 35 countries [8]. Each ERN works on a subset of RDs and maintains registries of varying complexity. Some ERNs have a single centralised registry to which participating healthcare providers submit data, whereas others have registries established in their participating institutes, where each institute collects and maintains its data.

Unfortunately, because each ERN collects unique data, there are wide variations in terms of content, format, and language across their RD registries. This heterogeneity makes it virtually impossible to jointly analyse ERN data, wasting considerable time and effort of data analysts and affecting any large-scale research project aimed at improving RD patient care. For instance, counts of patients with similar symptoms, treatments for similar symptoms across different geographic regions, or time-to-diagnosis cannot be produced by a simple query across all registries. A patient representative searching for "genomes pertaining to a rare disease profile not yet classified as such" or a researcher analysing "observed phenotypes of citizens with the same genetic profile" with the aim to "identify correlations with regional factors" are examples of more complex queries that can be executed on multiple resources across institutes and countries, the premises of which, however, is to make data Findable, Accessible, Interoperable, and Reusable (FAIR). It is, therefore, crucial to improve the Findability, Accessibility, Interoperability and Reusability (FAIRness or FAIR 'maturity') of the data collected in the RD registries of the 24 ERNs, for both humans and machines, as stated in the FAIR Guiding Principles [29]. When data are FAIR, they can be queried in an unambiguous and federated way, globally (if appropriate reuse conditions are met) without leaving its premises [30, 31]. In addition, an ecosystem based on FAIR principles adapts its functionality to its sources, because each source is self-explanatory.

Various methods can be applied for making data FAIR (also referred to as 'FAIRification') among the 24 ERNs, which contributes to diverging FAIRification methods and implementation choices throughout the network of ERNs. These differences are due to 1) different requirements and objectives (e.g., an initial focus on legal aspects, or a focus on internal queriability), 2) different software systems and tools (e.g., an Electronic Data Capture (EDC) system, lack of license for a specific ontology), 3) different disease domains (e.g., rare types of cancer, bone diseases), and 4) different jurisdictions (e.g., different laws between centres/countries). Applying different FAIRification methods theoretically still leads to interoperable solutions by definition, but overall, the process is not efficient for a community. Thus, harmonisation of methods and definitions and sharing of best practices would be beneficial to maximise the efficiency and benefit of FAIRification for all stakeholders.

Data can be made FAIR retrospectively, often long after they were collected, which may require extensive efforts to understand the meaning of the data [32, 33, 34]. Data can also be made FAIR when they are being collected, where the FAIRification steps are embedded in the data collection tool [35]. The latter was implemented for a VASCERN ERN registry, where data are made FAIR automatically and in real-time upon collection [36]. This FAIRification workflow can be reused by other ERNs across data collection platforms. Nevertheless, there is a need to guide the ERNs in achieving higher efficiency by aligning their implementation choices regarding tools (e.g., EDC software), standards (e.g., data representation syntaxes, ontologies), and legal decisions (e.g., sending data to a central registry in a different country versus several hospitals with their own FAIR databases, informed consent forms, data access policies, data processing and sharing agreements).

To harmonise FAIRification across ERN RD patient registries, a FAIRification steward team was established to act as liaisons between the ERNs and FAIR experts. These liaisons, supported by the EJP RD, provide a unique opportunity to investigate the ERNs' understanding and application of the FAIR principles to enable the use of data across international borders in the RD field. This work aims to 1) identify the challenges in FAIRifying RD registries and 2) support European-wide harmonised

FAIRification by proposing solutions in the RD field.

2.2 Methods

Organisation of the FAIRification Steward Team

The EJP RD FAIRification steward team was established on July 10th, 2020, to support and ensure harmonised FAIRification of ERN RD patient registries. The team is composed of six FAIR data stewards with different scientific backgrounds (biomedical science, software development, hospital management, public health, engineering) and education levels (BSc, MSc, and PhD). As illustrated in Figure 2.1, the FAIR data stewards facilitate the communication between ERNs and EJP RD FAIR experts. Each FAIR data steward collects FAIRification challenges from the ERNs they are assigned to. Then, the team curates these challenges and submits them to the FAIR experts, who provide the knowledge that is needed for proposing solutions. The team conveys the challenges requiring customised and ongoing support for a single ERN to the relevant experts and requests specific solutions.

Each ERN formed a core FAIRification team including a project manager or equivalent (e.g., data manager, registry manager), a clinical domain expert, a local data steward, and a developer. The last could be replaced by the hired EDC company programming support. Each FAIR data steward supports four ERNs and is the backup to four other ERNs. The communication channels between each ERN and their FAIR data steward were established in a first introduction meeting, and thereafter maintained in follow up meetings on demand.

Identification of the FAIRification Challenges

We identified the FAIRification challenges of the ERN RD patient registries in two main steps: collection of challenges and curation of challenges. The second step consists of three sub-steps: categorisation, rephrasing and merging of challenges. These are further detailed in this subsection.

Firstly, the FAIR data stewards collected the challenges that ERNs had with making their RD patient registries FAIR based on an initial set of 77 tools and standards identified by EJP RD FAIR experts. The implementation status of each standard or tool was identified for each ERN (“Implemented”, “Plans to Implement”, “Need Expert Help”, “Implementing Assisted by Expert” or “Non-Applicable”), as exemplified

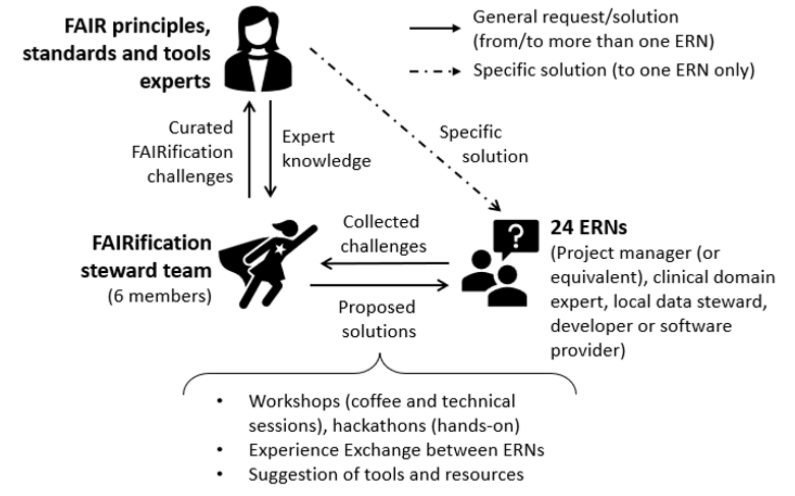

FAIR principles, standards and tools experts
General request/solution (from/to more than one ERN)
Specific solution (to one ERN only)
Curated FAIRification challenges
Expert knowledge
Specific solution
FAIRification steward team
(6 members)
Collected challenges
Proposed solutions
24 ERNs
(Project manager (or equivalent), clinical domain expert, local data steward, developer or software provider)
Workshops (coffee and technical sessions), hackathons (hands-on)
Experience Exchange between ERNs
Suggestion of tools and resources

Figure 2.1: FAIRification steward team, EJP RD FAIR (principles, standards, and tools) experts and European Reference Networks (ERNs) in a three-party interaction map. The FAIRification steward team works as liaisons between ERNs and EJP RD FAIR experts, collecting FAIRification challenges from ERNs, curating these challenges and providing them to experts, and returning consolidated knowledge from the experts to ERNs as proposed solutions. For single ERN requests, the team creates Expert-ERN communication channels (dashed line). The ERN team includes a project manager (or equivalent), a local data steward and a developer (or software provider). The set of proposed solutions comprehends workshops, where standards or tools are presented by experts; hackathons, where developers can try different tools themselves in a hands-on fashion; experience exchange between ERNs; and suggestions of existing implementations, tools, and resources.

in Table 2.1. Note that additional tools and standards could be added where applicable, as disclaimed in the document. Questions and implementation details specific to a tool or standard were recorded for each ERN and used as the main input for the FAIRification challenges. These data were collected by the FAIR data stewards while meeting with ERNs in a persistent and traceable document. To preserve privacy, access to this data is restricted to the associated EJP RD FAIR experts and FAIR data stewards. The FAIR data stewards continued to communicate with ERNs regularly to provide feedback and follow-up on their questions, which could lead to additional FAIRification challenges.

Secondly, all FAIRification challenges collected in the previous step by December 31st, 2020, were categorised, rephrased, and merged. All FAIRification challenges were categorised by: 1) “training”, specifying the need for training on a specific technology or concept; 2) “community”, requiring peer experience exchange; 3) “modelling”, re-

Function	**Tool/Standard Name**	**ERN Registry Implementation Status**
Data Model	CDE Semantic Model	*Implemented*
Set of data elements	Common Data Elements JRC	*Implemented*
Genes Ontology	HGNC	*Plans to Implement*
Genes Ontology	HUGO	*Non-Applicable*
Variant Ontology	HGVS	*Plans to Implement*
Phenotype Ontology	HPO	*Needs expert help (see methods)*
International Classification of Diseases	ICD-10	*Non-Applicable*
International Classification of Diseases	ICD-11	*Implemented*
Minimum Information About Biobank Data Sharing	MIABIS	*Implementing assisted by expert*

Table 2.1: An excerpt of the document used to collect the implementation status of each tool and standard for each ERN. The first column describes functions related to tools and standards which are listed in the second column. The last column tracks the implementation status of each tool or standard ("Implemented", "Plans to Implement", "Need Expert Help", "Implementing Assisted by Expert" or "Non-Applicable"). The references to the tools can be found in the template of the Additional file 1.

lating to (meta)data models or conceptual modelling activities; 4) "implementation", requiring programming expertise, such as the implementation of data exchange interfaces between systems; and 5) "legal", describing questions about data sharing and reuse agreements, informed consent, or any related services (e.g., patient informed consent form). These categories were defined by the FAIRification steward team based on the commonalities identified among the challenges. The categories and their definitions are summarised in Table 2.2. With this categorisation, we standardised the presentation of common solutions to avoid the need for repeated referrals to experts.

The FAIRification challenges after categorisation were rephrased and merged based on their content and commonalities. For instance, the two example challenges "We need hands-on help to implement the Common Data Element (CDE) [37] in REDCap (Research Electronic Data Capture) [38]" and "How can the CDE Semantic Model be implemented in Marvin XClinical [39]?" could be merged to one curated challenge "How to implement the CDE model [40] in my EDC system? ".

All processes, i.e., categorisation, rephrasing, and merging, were at least reviewed by two independent reviewers. The FAIRification challenges after this processing are referred to as curated FAIRification challenges. The remaining inconsistencies were resolved in discussions with the entire team and, upon need, with EJP RD FAIR experts.

Category	Definition
Training	Challenges related to inquiries for more information on a specific tool, standard, or a general concept.
Community	Challenges involving activities of peers in the same community to achieve reuse and prevent duplicated effort.
Modelling	Challenges involving the conceptualisation of data into data elements and bindings of standardised vocabularies to these data elements.
Implementation	Challenges involving implementation of a specific tool or standard
Legal	Challenges related to inquiries about data sharing and reuse agreements, informed consent, or implementation of related services.

Table 2.2: List of categories and their definitions. Five categories were created to organise the FAIRification challenges of RD patient registries. The categories reflect the nature of the challenges: the need for training, to learn from others, information about modelling, implementation, or legal aspects.

Proposing Solutions to the FAIRification Challenges

The FAIR data stewards defined solutions to the curated FAIRification challenges in collaboration with different stakeholders. The five stakeholder groups who contributed to the development of these solutions were: 1) ERN representatives, 2) EJP RD FAIR (principles, standards and/or tools) experts, 3) EJP RD coordinators, 4) Joint Research Centre, and 5) software developers and providers. To maximise efficiency, we defined solutions capable of addressing the highest number of challenges simultaneously. For the challenges that could be solved using readily available single solutions, we directly contacted the relevant stakeholders. Further, for the challenges that required novel solutions to be developed, the recombination of existing solutions, a long-term effort, or the participation of multiple parties, we arranged various types of activities that allowed for brainstorming for all stakeholders including ERNs.

2.3 Results

Here we present the work by the EJP RD FAIRification steward team to support the FAIRification of ERN RD patient registries. This includes the list of identified FAIRification challenges and proposed solutions to the ERNs. The solutions were reused or developed with input from multiple internal and external stakeholders to ensure convergence.

Overview of FAIRification Challenges

Ninety-eight FAIRification challenges were collected from all 24 ERNs. Their respective counts for each category before "original") and after curation are shown in Table 2.3. The most common category was "training" (31) while the least common was "community" (9). The "implementation" category contained 26 challenges, "legal" contained 20, and finally"modelling" contained 12. More details on all original and curated challenges can be found in the additional file [see Additional file 2].

FAIRification	**Categories**				
Challenges	**Train.**	**Comm.**	**Model.**	**Impl.**	**Legal**
Original (98)	31	9	12	26	20
Curated (41)	15	7	5	9	5

Table 2.3: The number of FAIRification challenges for each category (**train**ing, **comm**unity, **model**ling, **impl**ementation and **legal**) defined in our approach. The second and third rows show the number of challenges before and after curation, respectively.

After curation, the total number of challenges was reduced to 41. The "implementation" category had the biggest reduction (from 26 to 9). The "training" category was reduced from 31 to 15, "legal" from 20 to 5, "modelling" from 12 to 5, and "community" from 9 to 7. The "training" and "implementation" categories remained the most and second most common categories, respectively. On the other hand, "modelling" and "legal" were the categories with the lowest number of challenges after curation.

The fifteen curated "training" challenges were either related to a tool or standard, for example, CDEs, CDE semantic model [40], mapping languages, FAIR Data Point, registration of registries through European Rare Disease Registry Infrastructure (ERDRI) [41], informed consent, pseudonymisation, and query (see Table 2.4). "More information on semantic data model", and "More information on the FAIR Data Point (FDP)" are examples of "training" challenges.

The nine curated "implementation" challenges (see Table 2.5) were not only related to the tools and standards mentioned above but also "data format", to which 11 original challenges were merged. One example of these original challenges was "What are the recommendations for data formats for the EJP RD Virtual Platform?".

Curated Training FAIRification Challenges	Specific Solution
More information on ERDRI (added value, utility)	"Coffee rounds"" (ERDRI, Orphacodes, EUPID, Practical requirements, Practical implementation, Resource finder, Informed Consent, Disability and QoL)"
Documentation and specification of CDEs More information on CDE model (e.g. what is does, what is the added value, what would be the effort to implement it)	Documentation of Semantic CDE Model
More information on ADA-M and machine readable consent More information on Beacon 2.0 (added value, utility, how to implement it) More information on EJP-RD Metadata Model (what it does, what is the added value) More information on EUPID (e.g. licenses, costs) More information on FAIR Data Point (how will it work) More information on Phenopackets (utility) More information on Querying More information on RDF Mapping Languages What interoperability impact difference would be between using CDE Model and OMOP-CMD?	ERN Technical Workshops (Semantic CDE Model, EJP RD Metadata Model, EUPID API, Data formats and mapping languages, Phenopackets, Query builder, Orphacodes, DCDEs, PROMs)
More information on FAIR	Rome Summer School
Ground rules for interoperability (e.g., terminology, personnel, connectivity mechanism, API definition sets, diagrams, and technology specification) More information on EJP RD Virtual Platform	Virtual Platform Specification (VIPS)

Table 2.4: A summary of the identified training FAIRification challenges and proposed solutions. The first column lists the curated challenge, while the second describes the specific solution used to address that.

Curated Implementation FAIRification Challenges	Specific Solution
How can I use the iCRF generator tool? How to implement the CDE model in different EDC systems? Advice on data representation languages How to create RDF triples from a SQL database? How to integrate FDP in a registry? Is there a common template for excel import/exports (of the CDEs?)?	Experts from the CDE Modelling group for data conversion
Is there a template for batch import of metadata elements into ERDRI.MDR?	Experts from the EU RD Platform for findability of registries
How can the EJP RD metadata model be implemented?	Experts from the Metadata Modelling group for metadata conversion
How can the query builder tool be implemented on my system?	Experts from the Query Builder group for data querying

Table 2.5: A summary of the identified implementation FAIRification challenges and proposed solutions. The first column lists the curated challenge, while the second describes the specific solution used to address that.

The seven curated “community” challenges were related to the need for individual ERNs to learn from other ERNs (see Table 2.6). For instance, data sharing policies differ between healthcare providers at both the national and international levels, prompting ERNs to inquire about how the other ERNs dealt with such constraints.

The five curated “modelling” challenges (see Table 2.7) were all related to the CDEs but from different perspectives: how to interpret non-applicable CDEs (3); how to model non-compliant CDEs (3); how to interpret poorly defined CDEs (1); which ontology is recommended for a certain case (4); what if Orphanet is not sufficient for some RDs (1). For example, the data element “date of birth” is not allowed to be recorded due to national regulations, so only “birth year” is recorded. Another example is the WHO (World Health Organisation) Disability Assessment Schedule (WHODAS) [42]. It is a recommended standard for the data element “disability score”, but it does not apply to paediatric patients.

The five curated “legal” challenges (see Table 2.8) were related to legal concerns of the pseudonymisation tool (4) and its implementation (11), informed consent (1) and its machine-readable implementation (1), and data processing agreements and access policies (3).

Curated Community FAIRification Challenges	**Specific Solution**
How other ERNs annotate disability questionnaire?	Disability Survey
Exchange of experiences between ERNs registries What tools and standards do other ERNs use?	Exchange of FAIR experience
Learn from advanced registries with examples How do other ERNs share data?	Exchange of information on a regular basis
How do other ERNs collect the CDEs: 2.1 Date of Birth, 6.1 Diagnosis and 6.2 Genetic Diagnosis? What database templates do other ERNs use?	Share data dictionaries for DCDEs

Table 2.6: A summary of the identified community FAIRification challenges and proposed solutions. The first column lists the curated challenge, while the second describes the specific solution used to address that.

Curated Modelling FAIRification Challenges	**Specific Solution**
Which ontology is recommended for [X]?	CDE Modelling group
Are non-applicable CDEs mandatory? What if collected data do not follow the formats required in CDEs? What if the CDE [X] is not well-defined?	Experts from the JRC for CDEs
What if Orphacode is not sufficient for [X] diseases?	Experts from the Orphanet group for Orphacode

Table 2.7: A summary of the identified modelling FAIRification challenges and proposed solutions. The first column lists the curated challenge, while the second describes the specific solution used to address that.

Curated Legal FAIRification Challenges	Specific Solution
How can machine-readable information consent be modelled? Which consent form should be used?	EJP RD Consent Template
European level guidance on: Data Processing Agreements per database and countries; Agreements between EDC software and Hospitals that include multiple ERNs; ERNs Consortium agreement; Legal issues between countries.	ERICA project
How to implement EUPID within a registry? What are the legal concerns about the EUPID implementation?	Experts from the pseudonymisation tool

Table 2.8: A summary of the identified legal FAIRification challenges and proposed solutions. The first column lists the curated challenge, while the second describes the specific solution used to address that.

Overview of Proposed Solutions

Eleven types of solutions were proposed to address the different categories of FAIRification challenges (see Table 2.4). To address the “training” challenges, two types of solutions were proposed: 1) provide guidelines, and 2) organise training events. For the guidelines, EJP RD has created a list of deliverables [43] to establish concrete specifications that ERNs can adhere to. These deliverables include, for example, a report on the core set of unified FAIR data standards. For the training events, seven “coffee rounds” and eleven “technical workshops” [44] were organised. “Coffee rounds” were aimed to provide basic knowledge of tools or standards to a non-technical audience, whereas the “technical workshops” were designed to provide a more in-depth and technical understanding of how to implement a tool or standard. Through online surveys, the ERNs determined the topics and prioritised the order of the “coffee rounds” and “technical workshops”. The coffee round ”Introduction of the Orphanet nomenclature and the ORPHAcodes”, for example, introduced the concept of ORPHAcodes [45], clarified its objectives, and explained the benefits of its use. The “ORPHAcodes” technical workshop was organised to demonstrate how the standard could be implemented within an RD registry. Many of the “training” FAIRification challenges were addressed in the International Summer School on Rare Disease Registries and FAIRification of Data [46]. In this event, both FAIR data stewards and FAIR experts (EJP RD and external) were trainers.

Three solutions were proposed to address the "community" challenges: 1) survey ERNs and report on a specific challenge, 2) arrange experience exchange meetings, and 3) share information (see Table 2.6). In the first solution, a FAIR data steward got a request from their assigned ERNs on how peer ERNs resolved a particular challenge. For instance, WHODAS does not consider paediatric patients, which was insufficient to capture disability information in the domain of some ERNs. They then inquired whether other ERNs used alternative tools to assess the disability of paediatric patients in their registry. The FAIRification steward team then developed a survey on this request, which was disseminated to all ERNs by their assigned steward, respectively. The survey results were recorded and made available to ERNs upon request. The solutions were recorded to be used as input to the development of guidance tools.

In the second solution, the FAIR data stewards arranged experience exchange meetings between two ERNs when one ERN wanted to learn from (or collaborate with) another ERN at a more advanced stage in the FAIRification of their registry. Knowledge exchange between ERNs also contributes to the harmonisation of the FAIRification solutions across them. As an example, an exchange meeting was held between two closely collaborating ERNs that use the same platform and methods with common research interests in related diseases. This enables them to communicate with the FAIRification steward team as a single entity. Another example is an exchange meeting held between two advanced ERNs who wanted to exchange FAIRification experience and sought further collaboration regarding Patient-Reported Outcomes (PROs).

For the third solution, information sharing among ERNs was harmonised by FAIR data stewards. A typical example of this was that ERNs shared their data dictionaries (e.g., e-REC form in EuRRECa [47]) with the FAIRification steward team. Each ERN-specific data dictionary lists data elements to be collected in their registries together with definitions and accepted values.

The solution proposed to all "implementation" challenges is "to get implementation support from relevant experts", regardless of the tools or standards in question. The FAIR data stewards organised hackathons to define reference software implementations across ERNs (e.g., Implementation CDE Semantic Model for ERNs EDC providers [48]). These hackathons were held for individual ERNs, where FAIR experts gave hands-on support to a specific FAIRification challenge of an individual ERN.

Two types of solutions were proposed to address "modelling" challenges: 1) get modelling advice from relevant experts, and 2) organise a modelling group (see Table 2.7). The first solution mainly resolved challenges about ORPHAcodes [49] and CDEs, e.g., "how to model diseases that are not captured by ORPHAcodes", and "how do we interpret CDEs that are not well-defined". The second solution aimed to establish a dedicated modelling group for modelling discussions. The EJP RD CDE modelling group focuses on semantic data modelling (initially for CDEs, but now for other modelling needs) and provides support for addressing "modelling" challenges.

Three solutions were proposed to tackle challenges with informed consent, pseudonymisation, and data sharing policies in the "legal" category: 1) develop a generic consent form, 2) get implementation support from experts who develop the pseudonymisation tool, and 3) reach data processing agreements and data sharing policies (see Table 2.8). In the EJP RD, a generic consent form [50] involved European institutions. This generic consent form was subsequently translated into 25 national languages. The European Rare Disease Research Coordination and Support Action (ERICA) [51] Work Package 2 was created to support the ERNs in all the aspects related to data collection, integration and sharing, including ethical requirements.

2.4 Discussions

The proposed solutions to the FAIRification challenges presented in this paper contribute to increased harmonisation of FAIRification implementation decisions across ERN RD patient registries. Through workshops, the ERNs were not only connected to experts but also to other ERNs. These workshops created a collaborative environment for the exchange of ideas and implementation of solutions. The following subsection presents discussions on diversity in the FAIRification challenges, the strengths and weaknesses, lessons learned from the FAIRification steward work, and future work.

Diversity in FAIRification Challenges

The notable reduction in the total number of FAIRification challenges following curation from 98 to 41 (see Table 2) indicates that there are a considerable number of common ERN concerns (57), but also highlights the diversity among the challenges (41). The largest number (15) of curated "training" challenges reveals a gap in the knowledge and lack of access to training in the distinct aspects of FAIRification. Advice on data representation languages, the CDE semantic model, the EJP RD metadata model [52], mapping languages [53], and the pseudonymisation tool [54] are all examples of frequently encountered “training” challenges by ERNs. The other four categories are less diverse with their number of challenges ranging from 5 to 9, which becomes more evident in the “implementation” category, reducing from 26 to 9 curated challenges.

“Legal” challenges are mainly attributed to 1) the variation of legal documents required to collect, process and grant access to the data [55], and 2) the lack of awareness of EU-wide pseudonymisation tools. The variation of legal documents exists because of the country-specific legislation and different interpretations and applications of GDPR (General Data Protection Regulation) [56, 57]. Some countries even request additional safeguards for sensitive data, which increases the complexity of establishing a patient registry. The lack of awareness of using EU-wide pseudonymisation tools was another practical issue that resulted in some of the “legal” challenges. Given that some ERNs already had an internal pseudonymisation system in place, they questioned the added value of using an additional pseudonymisation tool and were concerned about the cost of re-assigning pseudonyms to existing patient records. Currently, the European Joint Research Centre (JRC) is working on the development of an EU-wide pseudonymisation tool to be reused by the ERNs.

The “community” challenges refer to how others use standards and tools. The fact that ERNs look for reusing peer solutions fosters convergence and interoperability. This is a positive observation because standards are only interoperable when used across organisations [58]. In fact, the third foundational principle of FAIR, *Interoperability*, is the most challenging one to be realised [59], and consequently, requires considerable effort [60]. Once the community standards are agreed upon, the reusability of data is facilitated, contributing to a sustainable scientific environment [61]. Convergence over the tools and standards used to promote interoperability within the community is necessary and will benefit new registries in general. Thus, it enables the RD community to define its FAIR Implementation Profile [61], a list of community-supported choices that promote convergence for FAIRification. In general, interoperability issues extended beyond technical FAIR standards to include legal and modelling considerations. For instance, country-specific legislation may prohibit the collection of certain data elements, thereby directly impacting modelling and thus the data sharing capabilities between registries from different countries. To support these legal and ethical challenges, EJP RD offers a helpdesk and an AREB (Advisory Regulatory Ethics Board) [62] office.

Strengths and Weaknesses of the approach

By forming a team of FAIR data stewards from diverse backgrounds we were able to harmonise the disparate FAIRification procedures of RD registries. The workload was balanced efficiently among the stewards enabling effective communication with ERNs. This consulting experience resulted in increased networking, convergence, and dissemination of knowledge. Besides, the FAIRification challenges of the ERNs were gleaned as first-hand information by the FAIR data stewards. Therefore, the challenges could accurately reflect the actual issues faced by RD registries in their EU-wide FAIRification and serve as valuable information for decision-making at the project level.

When compared to our previous FAIRification experience involving a single FAIR data steward [35], a team supported FAIRification effort resulted in a more robust approach. First, the diverse backgrounds and the collaboration among team members facilitated experience exchange and FAIRification discussions. This has enabled the stewards to scrutinise FAIRification challenges from a variety of angles, resulting in the development of a collection of diverse solutions. Secondly, such team-based support enables the stewards to maintain a consistent pace in the communications with the ERNs, for example by the support of backup stewards, as one person could assist an overwhelmed teammate when necessary.

Since each of the FAIR data stewards may have had slightly different discussions in their regular meetings with the ERNs, there may have been differences in the way each ERN described their challenges. However, this bias was reduced by using the initial set of tools and standards (see the Methods Section) as a starting point for these discussions. The same is true for the interpretation of the original FAIRification challenges by each of the FAIR data stewards. The rephrasing style may have varied and influenced the final number of curated challenges. To mitigate this problem, we performed cross-checking between pairs of stewards, so one could validate the rephrasing and merging of the other.

Nonetheless, the significance and implications of our findings, particularly in the progress of RD registries FAIRification, reinforce the importance of this type of work. The steps taken by the FAIRification steward team to communicate, collect information and identify solutions can, therefore, be reused in guidance to other FAIR project management in general. In addition, the sustainability of any approach developed in the EJP RD is a core value of the project that also concerns the FAIRification steward team. In September 2021 during the EJP RD general assembly, a workshop [63] on the sustainability of the FAIRification steward service was held, and it was concluded that this EU-wide service should be continued and made available to other types of resources apart from registries.

Lessons learned

The unique experiences from the interaction between the FAIR data stewards and the diverse RD registries are summarised below:

- *Clarify FAIRification goals before implementation.* Available FAIRification workflows recommend that defining the FAIRification goal(s) is the first key step in FAIRification [32, 35]. Nonetheless, some of the RD registries did not complete this step yet. When defining clear FAIRification objectives, the local FAIRification team will be able to make smarter choices that are aligned with the goals. Additionally, by understanding the FAIRification context and aims, the team can be more motivated to go through the implementation process.

- *Have access to FAIR experts.* FAIRification knowledge is complex and multifaceted, which raises the need to establish the connections of standards and tools with the FAIRification workflow. For that purpose, a network of experts specialised in FAIRification of research data is needed. Access to such a wealth of expertise also aided the FAIR data stewards in the development of guidelines.

- *Attend active training about FAIR(ification).* While collecting the FAIRification challenges, we realised that there was a significant difference in the perception of FAIR between the different ERNs. Some were unfamiliar with the FAIR principles while others had different interpretations of them. As a result of this knowledge gap, some ERNs may have faced similar challenges but articulated them differently. To reach a consensus on FAIR literacy as well as FAIR awareness, attending workshops and hackathons to share experiences and brainstorm ideas is of foremost importance.

- *Use the Common Data Elements (CDEs) and their semantic data model.* Collecting the CDEs can increase interoperability among ERNs, but, even if a prespecified list of CDEs is provided, there are still many challenges regarding compliance and interpretation with that list. Thus, representing these CDEs through a semantic data model in a machine-readable fashion is needed to reduce ambiguity. Further, since the CDEs do not capture various domain-specific data elements, a new list of Domain-specific Common Data Elements (DCDEs) is being developed by the EJP RD to be applied to the RD registries.

- *Have vendors incorporate FAIRification in the data collection software.* ERN registries are dependent on various software to collect and manage their data. When a FAIRification workflow is embedded in the registry data collection process, the burden of making data FAIR is reduced.

- *Define and reuse community standards.* Standards and implementation choices should be defined and reused within the related research community to converge and harmonise by default.

- *Resolve legal issues internationally in a FAIR optimised way.* By legal issues, we refer to pseudonymisation, informed consent, data processing agreements and data sharing policies. Any disagreements between these can become the bottleneck that hinders many steps of FAIRification and drags out the entire process. In addition, tackling these issues is time-consuming and labour intensive but still necessary, which requires dedicated negotiations across countries.

Future work

At this date, large efforts have been deployed to support ERNs with the CDEs implementation and FAIRification. In the second year of work, the FAIR data stewards collected and compared ERNs' data dictionaries to identify common research, disease, or domain-specific data elements (DCDEs). The goal of DCDEs is to reach convergence and standardisation of what and how ERNs collect data elements other than CDEs, thereby increasing interoperability, facilitating collaborative research, and improving data discoverability. An additional advantage is that the newly identified commonly collected data elements will be semantically modelled by EJP RD experts in close collaboration with the domain experts who are choosing the DCDEs. Separating these processes and the modellers from the domain experts, as was the case for the CDEs, makes accurate modelling much harder. They are also expected to be added to ERDRI to encourage reuse by new RD registries.

The challenges presented in this study were collected as one of the FAIR data stewards' initial tasks. In the future, we plan to further support FAIRification, by providing a Smart Guidance tool. This tool will combine the knowledge, results, and resources of the FAIR data stewards and EJP RD FAIR, and create an interactive questionnaire that generates a personalised FAIRification plan. A partial preview of the

Smart Guidance content can be found in the visual representation called FAIRopoly [64].

The FAIR data stewards will continue to support the FAIRification of ERN registries. We will first reassess the implementation status of standards and tools used in the ERNs registries FAIRification to learn the effect of our FAIR guidance and proposed solutions. We also plan to support the FAIRification of other EJP RD resources.

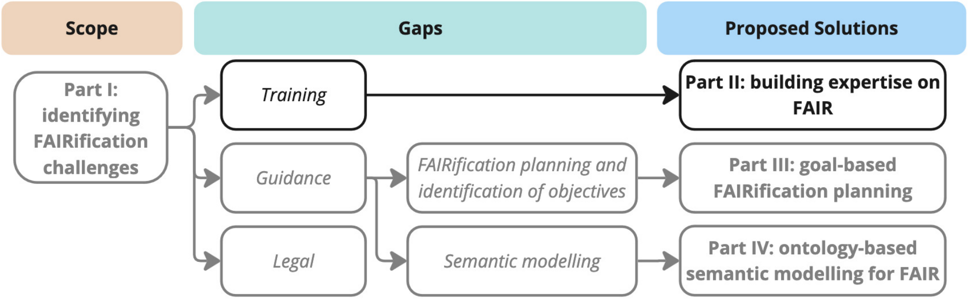
2.5 Conclusion

We identified the main challenges faced by RD registries during FAIRification and proposed collaborative solutions to address them. ERNs desire to learn about EJP RD-recommended tools and standards for facilitating FAIRification, and have a high demand for assistance in implementing these tools and standards. This overview is a valuable resource for EU-wide FAIRification efforts in the RD field. For example, the most common challenge may be the most significant bottleneck, and therefore might be prioritised in most FAIRification initiatives.

As FAIR data stewards, we supported the harmonisation of solutions for making RD data FAIR across countries, and continue to futnction sustainably, as motivated by the work described in this paper. We anticipate that our findings and lessons learned will increase FAIR awareness in the RD field and provide suggestions for other large FAIRification efforts. Specifically, we foresee that our unique team-based setup for supporting FAIRification will be adopted by other projects to recreate a similar hovering consultant team.

Part II

Building expertise on FAIR


Scope
Gaps
Proposed Solutions
Part I: identifying FAIRification challenges
Training
Guidance
Legal
FAIRification planning and identification of objectives
Semantic modelling
Part II: building expertise on FAIR
Part III: goal-based FAIRification planning
Part IV: ontology-based semantic modelling for FAIR

Topics covered in Part II of this thesis.

This part tackles the increasing need of expertise on FAIR, which is one of the FAIRification challenges identified in the previous part. For this, Chapter 3 describes the Bring Your Own Data (BYOD) workshops to help researchers make their data FAIR, and the three BYOD formats designed for training different types of stakeholders: the data-focused BYODs educate domain experts on how to make their data FAIR, the management-focused BYODs instruct managers on the benefits and characteristics of making data FAIR, and the software-focused BYODs gather software developers and experts on FAIR to implement or improve other resources that are used in FAIRification. Participants in the BYODs learn about FAIR principles and their benefits through hands-on experiences.

Chapter 3

Building expertise on FAIR through evolving Bring Your Own Data (BYOD) workshops

Describing the data, software, and management focused approaches and their evolution

César H. Bernabé*, Lieze Thielemans*, Rajaram Kaliyaperumal*, Claudio Carta*, Shuxin Zhang, Celia W.G. van Gelder, Nirupama Benis, Luiz Olavo Bonino da Silva Santos, Ronald Cornet, Bruna dos Santos Vieira, Nawel Lalout, Ines Henriques, Alberto Cámara Ballesteros, Kees Burger, Martijn G. Kersloot, Friederike Ehrhart, Esther van Enckevort, Chris T. Evelo, Alasdair J. G. Gray, Marc Hanauer, Kristina Hettne, Joep de Ligt, Arnaldo Pereira, Núria Queralt-Rosinach, Erik Schultes, Domenica Taruscio, Andra Waagmeester, Mark D. Wilkinson, Egon L. Willighagen, Mascha Jansen, Barend Mons, Marco Roos, Annika Jacobsen

**shared first authorship*

Data Intelligence Journal, 6, 2 (2024). https://doi.org/10.1162/dint_a_00236

Abstract

Since 2014, "Bring Your Own Data" workshops (BYODs) have been organised to inform people about the process and benefits of making resources Findable, Accessible, Interoperable, and Reusable (FAIR, and the FAIRification process). The BYOD workshops' content and format differ depending on their goal, context, and the background and needs of participants. Data-focused BYODs educate domain experts on how to make their data FAIR to find new answers to research questions. Management-focused BYODs promote the benefits of making data FAIR and instruct project managers and policy-makers on the characteristics of FAIRification projects. Software-focused BYODs gather software developers and experts on FAIR to implement or improve software resources that are used to support FAIRification. Overall, these BYODs intend to foster collaboration between different types of stakeholders involved in data management, curation, and reuse (e.g. domain experts, trainers, developers, data owners, data analysts, FAIR experts). The BYODs also serve as an opportunity to learn what kind of support for FAIRification is needed from different communities and to develop teaching materials based on practical examples and experience. In this paper, we detail the three different structures of the BYODs and describe examples of early BYODs related to plant breeding data, and rare disease registries and biobanks, which have shaped the structure of the workshops. We discuss the latest insights into making BYODs more productive by leveraging our almost ten years of training experience in these workshops, including successes and encountered challenges. Finally, we examine how the participants' feedback has motivated the research on FAIR, including the development of workflows and software.

3.1 Introduction

The FAIR data principles address critical factors to make the analysis of multiple sources more efficient by improving their Findability, Accessibility, Interoperability, and Reusability for humans and computers [1]. The process of making data FAIR ("FAIRification"), although partially supported by software (e.g. data transformation tools) and standards (e.g. ontologies), relies on expert knowledge about the data generating domain (domain experts and data owners) and about FAIR-related aspects such as metadata design, conceptual modelling, licensing definition, and use of identifiers (FAIR experts). Given the high demand to make resources FAIR and a shortage of FAIR expertise, a series of "Bring Your Own Data" workshops (BYODs) have been organised since 2014 to bring together expert knowledge to accelerate the practical FAIRification of resources (e.g. datasets, registry information, ontologies). The bidirectional learning experience between domain and FAIR experts results in making the BYODs a mutually beneficial experience. Attendees (domain experts) receive hands-on guidance in making their data FAIR, while trainers (FAIR experts) gain valuable insights to improve their own training skills, topics, and materials, and develop more effective FAIR support tools, processes, and guidelines.

The first BYOD workshop took place six months after the "Jointly Designing a Data FAIRPORT" workshop [65], which marked the inception of the FAIR principles. Subsequently, the first BYOD for rare disease registries and biobanks, held in November 2014, initiated an annual series of BYODs for this specific domain. Initially, data-focused BYODs emerged from the need to train people on FAIRification and therefore focused on making data resources FAIR. Over time, as the FAIR community matured, it became clear that different types of BYODs were necessary to meet different contexts, needs and backgrounds of attendees. Consequently, two additional types of BYOD structures were designed: management- and software-focused BYODs. The former aims at informing managers and policy-makers on the added benefits of FAIR and the requirements for FAIRification. The latter focus on developing software tools that support the process of making data FAIR, or tools and standards that enhance the FAIR level of resources.

The remainder of this paper is organised into four sections. The next section describes the three types of BYODs. The section "the evolution of BYODs" lists the BYODs organised since 2014 and reports on the first BYOD workshop and the series of BYODs for rare disease registries, with emphasis on how these workshops have led to the improvement of the content and didactical aspects of the BYODs. Then, we discuss the impact of the BYODs on the FAIR community and on the expert domains (i.e. plant breeding, and rare disease registries and biobanks). In this paper, we mention different types of experts. For clarity, we refer to "X expert" as an expert specialised in a certain tool, standard, or knowledge. For instance, "FAIR expert" refers to an expert with knowledge and experience in FAIR.

3.2 The three types of BYOD structure

Despite catering for different contexts and types of participants, all BYODs share the same overarching goal of fostering expertise in FAIR-related topics while cultivating community confidence in the benefits of having FAIR resources. Table 3.1 summarises the main aspects of the three different BYOD structures, which are further described in the subsections that follow.

	Data	**Management**	**Software**
Required technical knowledge for attendees	Intermediate	Low	High
Learning Format	- Knowledge exchange - Hackathon - Lectures	- Knowledge exchange - Lectures	- Knowledge exchange - Hackathon
Main goals	- Answer research questions using FAIR data - Make resources FAIR - Train domain experts on FAIR(ification)	- Inform participants about the added benefits of FAIR - Inform managers on the characteristics of FAIRification projects	- Develop software that support enabling FAIR on resources (e.g. FAIR Data Point) - Develop resources that support the FAIRification process (e.g. data transformation tools)
Main tasks	- Semantic modelling of (meta)data - Hosting and querying of FAIR data	- Hands-on scenarios simulating the benefit of FAIR resources - Plenary sessions to discuss the characteristics of FAIRification projects	- Solution brainstorming - Solution implementation - Testing of implemented solution
Profile of attendees	- Researchers - Domain experts	- Project managers - Registry managers - Patient representatives - Policy makers - Funders	- Researchers on FAIR - Software Developers - FAIR data stewards - Domain experts

Profile of trainers	- Experts in FAIR - FAIR data stewards - Ontologists - Standards specialists	- FAIRification project managers - FAIR data stewards - Decision/policy makers with knowledge on FAIR	- Researchers on FAIR - Developers - FAIR data stewards - Domain experts
Commonly used tools and standards	- Ontologies and metadata models - FAIR Data Point - Domain specific standards (e.g. CDE Semantic Model [ref]) - FAIR enabling standards (e.g. DCAT [40])	- FAIRification workflows - Collaborative brainstorming tools (e.g. mind maps, black boards)	- Collaborative brainstorming tools - Software development resources (e.g. programming languages such as Python)
Expected Outputs	- FAIR resource - Answer to research question(s)	Audience knowledgeable about: - the characteristics of FAIRification - the added benefits of FAIR	- FAIR enabling resource - FAIRification supporting resource

In addition to the two- or three-day duration of the BYODs, some also contain preparatory and follow-up phases for attendees and trainers. Attendees are invited to participate in introductory webinars, to prepare for the workshop, and post-BYOD follow-up meetings, for support on subsequent activities. The introductory webinars aim to familiarise attendees with FAIR and initial FAIRification needs (e.g. identification of goals and required domain expertise). In the post-BYOD follow-up meetings, attendees are advised on other FAIRification challenges that might appear after the BYOD. The preparatory and follow-up phases provide opportunities for trainers to prepare and evaluate the workshop agenda, training materials, and methods of instruction. Follow-up phases are also used by trainers to plan for improvements in future editions of the workshop in response to feedback from participants.

The three different BYOD formats herein described are intended to guide other institutions and groups in organising their own BYODs. These formats can be freely adapted by any community to suit their own learning goals, needs and constraints. We suggest that BYODs are organised with a multidisciplinary training group, including at least a FAIR expert, a conceptual modelling expert and an expert in the domain of the resource to be made FAIR.

3.2.1 Data-focused BYOD workshops

During the data-focused BYODs, attendees are divided into groups, with at least one trainer allocated per group. Each group can use their own data or request synthetic data. Collaboratively, the groups transform their data into FAIR data by following a step-by-step FAIRification process.

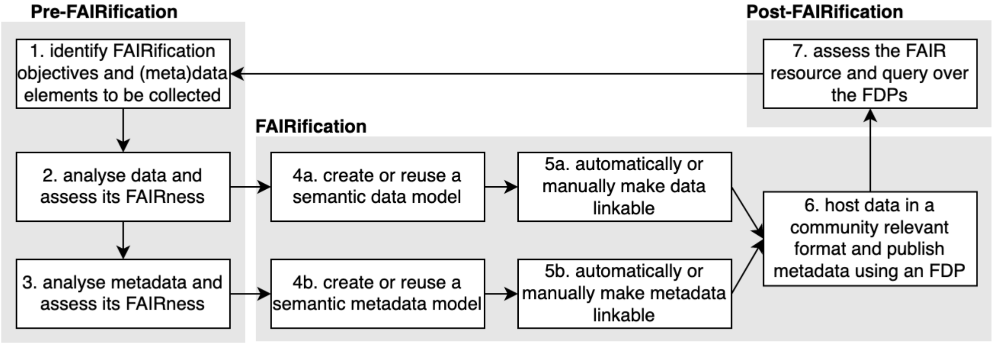

Pre-FAIRification
1. identify FAIRification objectives and (meta)data elements to be collected
2. analyse data and assess its FAIRness
3. analyse metadata and assess its FAIRness
FAIRification
4a. create or reuse a semantic data model
4b. create or reuse a semantic metadata model
5a. automatically or manually make data linkable
5b. automatically or manually make metadata linkable
6. host data in a community relevant format and publish metadata using an FDP
Post-FAIRification
7. assess the FAIR resource and query over the FDPs

Figure 3.1: Illustration of the FAIRification workflow used in the data-focused BYODs, modified from [2] and [35]

In most workshops, we have followed a FAIRification workflow where data is made FAIR retrospectively (after data collection - *post hoc*) and semi-automatically (see Figure 3.1). This workflow was developed based on emerging FAIRification steps observed in early BYODs. It should be noted, however, that previous BYODs may have deviated from this structure while evolving towards the current format. Additionally, more recent workshops focused on making data FAIR by design (automatically during data collection - *de novo*) (e.g. [35]). Figure 3.1, which is adapted from [2] and [35], illustrates the FAIRification workflow used in recent BYODs. The workflow is divided into three phases: 1) Pre-FAIRification, 2) FAIRification, and 3) Post-FAIRification, which are each subdivided into clear steps. These hands-on phases are usually accompanied by lectures about FAIR related topics and plenary sessions where participants can share their experiences with FAIRification, including challenges and success cases (as listed in Table 3.1).

The Pre-FAIRification phase is composed of three steps. In step 1, to identify FAIRification objectives and (meta)data elements to be collected, the groups define driving objectives and research question(s) focusing on using their sample data in combination with other FAIR data. Next, following steps 2 and 3, the groups closely investigate the representation (syntax) and meaning (semantics) of their data and the metadata (i.e. description of data). In our experience, metadata such as the (meta)data's licence and provenance information is often not available *a priori* and therefore needs to be gathered during the BYOD workshop. Finally, before doing any actual FAIRification, the FAIR status of the data is assessed by using tools such as the FAIR Evaluation Services [66] (step 3) (see [67] for other FAIR assessment services).

In the FAIRification phase, the groups create or reuse a conceptual model to describe the data elements and their relationship (step 4a), and a metadata model to provide information about the data (step 4b). These conceptual models must contain, at a minimum, the data elements required to answer their driving research question(s). These conceptual models are semantically enriched by binding the models' concepts to terms from reference ontologies. In step 5a, the data is made machine-readable (i.e. in a format that can be processed by a computer) by using the semantic data model and existing tooling to generate an ontologised version of the data manually (e.g. FAIRifier [68]) or automatically (e.g. Castor EDC [6], MOLGENIS [69]). The metadata is also made machine-readable (step 5b) by using metadata standards (e.g. Data Catalogue Vocabulary (DCAT) [70]). Finally, the machine-readable metadata is made available using the FAIR Data Point (FDP) (step 6) [71] and the machine-readable data is hosted using a community relevant file format (e.g. RDF [72]).

In the post-FAIRification phase, the driving research question(s) defined in step 1 are answered using the newly created FAIR data (step 7). Here, the FAIR status of the data resource is reassessed and compared to the assessment done in the pre-FAIRification phase to verify if the improvement of the FAIR level of the data meets the goals previously defined.

To illustrate, in the pre-FAIRification phase, a group defines "finding new treatment candidates for untreated rare disease patients" as a driving goal, and reuses data from different rare diseases registries to achieve this goal. The data includes information on diagnosis, symptoms and treatments, and metadata includes information such as the (meta)data's licence (e.g. CC BY-NC 4.0 [73]) and provenance (e.g. from which registry the data originates). In the FAIRification phase, the group adopts the CDE Semantic Model [20] (step 4a) as the semantic data model and the EJP RD metadata model [74] as the semantic metadata model (step 4b). Reusing the ontologies from the models adopted in steps 4a and 4b supports making the data and metadata linkable (steps 5a and 5b). Finally, the newly linked (meta)data is hosted and published using a FDP (step 6). During the post-FAIRification phase, the group leverages the FAIR data they have created to address their research question by writing federated queries. For instance, they may query their FAIR data with other public resources to identify treatment candidates for patients with similar symptoms.

3.2.2 Management-focused BYOD workshops

The management-focused BYOD workshops are geared towards informing registry and project managers, patient representatives, and decision-makers about the characteristics of FAIRification, including the associated costs, time, expertise, and effort required. The need for this type of BYOD emerged due to the growing adoption of FAIR in various institutions, which has required personnel in high-level positions to become familiar with the benefits and prerequisites of data FAIRification. As a result, these workshops place less emphasis on technical work and more on general considerations of FAIRification. The management-focused BYOD is divided into three phases: (i) understanding the problem of not having FAIR data, (ii) acquiring knowledge about FAIR and FAIRification, and (iii) training on FAIRification project management.

The first phase of a management-focused BYOD is executed in an interactive manner, typically through the use of simulated case scenarios that recreate the challenge of dealing with incomprehensible and non-interoperable data. To highlight the importance of FAIR data, attendees are tasked with challenges that require connecting data from multiple sources, while being presented with non-standardised and multilingually annotated data in different formats, making the task more difficult to accomplish.

In the second phase, attendees learn about the benefits of FAIR and the main steps of FAIRification. The learned benefits aim to address the challenges identified in phase one. Plenary and hands-on sessions provide practical experience in FAIRification related tasks, including conceptual modelling, making metadata findable, and using ontologies and FAIR compliant Electronic Data Capture (EDC) systems. This phase is typically concluded by revisiting the mock case from the first phase, but this time using FAIR data, thus demonstrating how the previously identified challenges can now be solved more easily and efficiently.

In the third phase, participants discuss the implications (e.g. budget, time, required expertise and infrastructure) of FAIR for project managers and policymakers. After the plenary sessions, attendees have a hands-on session on how to create their FAIRification team.

A real-world example of this structure can be visualised on the agenda of recent management-focused BYODs organised for rare disease registries and biobanks (e.g. [75]). For instance, in the one held online in 2022 [75], attendees experienced the problems of not having FAIR data through a digital game where they had to find treatments for new patients in different datasets organised in a non standardised manner (e.g. using synonyms for equivalent concepts) and presented in several languages (e.g. Mandarin, Dutch, and Spanish). Thereafter, lectures and discussion sessions on topics such as FAIRification steps, conceptual modelling, ontologies and querying informed the attendees about FAIR-related aspects. On the second day, the attendees played the same digital game, only this time with FAIR data, which allowed them to accomplish the goal of finding treatment in distributed datasets. After lectures and discussions on the benefits of FAIR, participants exchanged experiences about the implications of data FAIRification for registry managers.

3.2.3 Software-focused BYOD workshops

The main goal of software-focused BYOD workshops is to create software that supports FAIRification, or software and standards that increase the FAIR level of resources, as shown in Table 3.1. Participants of software-focused BYODs include researchers working on FAIR-related projects, FAIR data stewards [76], developers and, in certain cases, domain experts. In this type of workshop, trainers and attendees come from similar backgrounds, working together to exchange knowledge while tackling the same goal. This type of BYOD is organised in a hackathon setting with five phases:

1. Understanding the problem: participants discuss the need or problem to be solved (e.g. facilitating metadata publication)

2. Proposing solutions: participants are invited to brainstorm solutions (e.g. using brainstorming tools such as mind maps) to the problem described in the previous phase (e.g. developing software to support the creation and publication of FAIR metadata)

3. Prioritising tasks: the implementation tasks are ordered by importance, and then selected for implementation (e.g. developing a proof-of-concept software that creates machine-readable metadata from an Excel sheet and publishes it via an FDP)

4. Coding and experimenting: the prioritised tasks are implemented and the resulting implementation is tested (e.g. implementing and testing the proof-of-concept software with mock data).

5. Reporting: the implementation is reported and published (e.g. a paper or website documenting the script developed during the hackathon)

Most software-focused BYODs are typically structured around iterative cycles. After a set period of time, participants convene to report on their group status and the findings from their tasks. They can then decide to switch or merge groups, get advice from others and/or continue their tasks. As a result, the agenda of the software-focused BYOD is adaptable to the requirements that emerge during the workshop. At the end of the BYOD, conclusions on which solutions to follow up are made. Adaptations of current tools, prototypes, proof-of-concept implementations or architectural designs are examples of outcomes of software-focused BYODs.

The "hackathon to make MOLGENIS FAIR" [77], which took place in 2016, is a real-world example of a software-focused BYOD. During this event, software developers and FAIR experts worked collaboratively to create a proof-of-concept of making MOLGENIS FAIR. MOLGENIS [69] is an open-source data platform for the management of scientific data. By the end of the hackathon, the team had implemented an application programming interface (API) to publish datasets in MOLGENIS as FDPs.

3.3 The evolution of BYODs

Table 3.2 highlights the BYODs held from 2014 to 2023. For context, the table includes the date of the "Jointly Designing a Data FAIRPORT" Workshop, where the FAIR principles were initially conceived, and the publication of the paper describing the FAIR principles [1]. A list with more detailed information on the workshops is available as supporting information.[1]

BYODs listed by Table 3.2 include the first workshop on genetic biodiversity [78] and the Bring Your Own Rett Syndrome Data workshop [79]. The former was a data-focused BYOD where participants worked on linking different datasets (e.g., the Centre for Genetic Resources (CGN) tomato collection and phenotypic observations, and variants from the 150 tomato genome re-sequencing project [80]) that were then queried as combined data. The latter focused on producing FAIR nanopublications[2] about Rett Syndrome, the results of which led to an ELIXIR implementation study on

[1] `https://doi.org/10.5281/zenodo.8155154`

[2] Nanopublications are defined as the smallest publishable unit of facts with full information where the knowledge comes from.

the interoperability of molecular data in rare diseases (MolData2) [81] and contributed to the development of the cross-omics data analysis work package of the EJP RD.

Title	Date and Location	Focus
FAIR Principles Idealisation - Jointly Designing a Data FAIRPORT	13 - 16 Jan 2014 - Leiden, the Netherlands	-
The first BYOD workshop	24 - 25 Jun 2014 - Leiden, the Netherlands	Data
The first Bring Your Own Data (BYOD) Workshop To Link Rare Disease Registries - First RD BYOD Workshop	26 - 27 Nov 2014 - Rome, Italy	Data
The first ”green genetics” BYOD	21 - 22 Jan 2015 - Wageningen, the Netherlands	Data
The Bring Your Own Template (BYOT) workshop	20 Feb 2015 - Utrecht, the Netherlands	Data
The second RD BYOD Workshop	24 - 25 Sep 2015 - Rome, Italy	Management
FAIR Principles Paper Published	15 Mar 2016	-
The third RD BYOD Workshop	29 - 30 Sep 2016 - Rome, Italy	Data
The FAIR Data Hackathon	19 - 20 Oct 2016 - Utrecht, the Netherlands	Software
The Software Solution Provider BYOD	25 - 27 Oct 2016 - Leiden, the Netherlands	Software
The Bring Your Own Rett Syndrome Data workshop	1 - 3 Nov 2016 - Maastricht, the Netherlands	Data
How to Make Data FAIR for Open Science	15 - 19 May 2017 - Leiden, the Netherlands	Data and management
The plant phenotype BYOD and hackathon	30 May - 1 Jun 2017 - Ghent, Belgium	Software
The cancer genomics BYOD	6 - 8 Jun 2017 - Utrecht, the Netherlands	Data and software
The fourth RD BYOD workshop	21 - 22 Sep 2017 - Rome, Italy	Data
The DSM BYOD workshop	25 - 26 Sep 2017 - Delft, the Netherlands	Data
The fifth RD BYOD workshop	13 - 14 Sep 2018 - Rome, Italy	Data and management
The RIKILT/WUR BYOD workshop	22 Nov 2018 - Wageningen, the Netherlands	Data
BYOD FAIRification workshop at Leiden University Library	18 Jun - 2019 - Leiden, the Netherlands	Data
The sixth RD BYOD workshop	26 - 27 Sep 2019 - Rome, Italy	Data and management
The seventh RD BYOD workshop	1 - 2 Oct 2020 - Online	Management
The eighth RD BYOD workshop	30 Sep - 1 Oct 2021 - Online	Management
The ninth RD BYOD workshop	29 - 30 Sep 2022 - Online	Management

The World Duchenne Organization's FAIR Training Program	7 - 9 March 2023 - Online	Management
The tenth RD BYOD workshop	28 - 29 Sep 2023 - Rome, Italy	Management

Table 3.2: Overview of 'Bring Your Own Data' (BYOD) workshops organised from 2014 to September 2023.

All BYODs have played an important role in iteratively improving the structure of subsequent ones, as well as in facilitating the adoption and research on FAIR. As representative examples, the following subsections describe the inaugural BYOD – the Human Protein Atlas and MycoBase BYOD – and the series of BYOD workshops focused on linking rare disease registries. The BYODs evolved based on the trainers' perception during the workshops and based on informal feedback from attendees. No direct data from any participant was collected.

3.3.1 The first BYOD workshop: Human Protein Atlas and MycoBase

The first data-focused BYOD workshop was held in Leiden, the Netherlands, on the 24th and 25th of June, 2014 [82, 83]. It was organised by a group of researchers from across Europe, and sponsored by the Dutch Techcentre for Life Sciences (DTL) [84], and Elixir [85]. The BYOD, which focused on making data interoperable, brought together data owners from the Human Protein Atlas [86] and MycoBase [87] with Linked Data experts. It is important to note that the FAIR principles and, consequently, the concept of "FAIR" data were still under development by then. Therefore, this BYOD focused on creating "Linked Data", which is a step towards having FAIR data.

The data owners needed to be familiar with their current internal data management structures, i.e. the database schema and data pipelines for creating and displaying entries. The Linked Data experts had a variety of backgrounds such as semantic web services [88] and integration platforms [89]. The main aim was to develop sample Linked Data to demonstrate the added value of interoperable data for facilitating answering research questions by reusing information from multiple resources.

The BYOD event started with a plenary training session providing an overview of Linked Data. Then, the attendees were split into two working groups, each of which aimed to develop a proof of concept centred around one of the data resources, driven by their own research questions. The Human Protein Atlas group focused on developing a subset of Linked Data from the Human Protein Atlas. This was then connected to WikiPathways data [90]. The MycoBase group linked their data with the content of ChEMBL through the Open PHACTS API [91, 92]. The Human Protein Atlas developers have used the experience of the event to develop their own RDF data release, heavily reusing the ontological model of neXtProt [93].

In this BYOD, it became clear there was a need to include preparation and follow-up meetings in the agenda of subsequent workshops. The experience gained by organisers provided insights for planning pre-BYOD training about the FAIR principles and for organising post-BYOD supporting sessions. Additionally, it highlighted the importance of publishing training materials that could be used at other BYOD events, promoting knowledge sharing and dissemination.

3.3.2 A series of annually recurring BYODs: rare diseases registries

Making rare disease resources interoperable and, thereby, preparing them for multi-source analysis is crucial since rare diseases occur at low frequency. In Europe, a disease is considered rare when it affects less than 5 in 10,000 individuals [94]. Ensuring the interoperability of rare diseases data is important because non-integrated data would likely be insufficient to support research or improvements in outpatient care. Therefore, each local data resource is of relatively limited value on its own, but may be highly valuable in combination with other data.

The first BYOD for rare disease registries and biobanks[3] (RD-BYOD) was held in Rome, Italy, at Istituto Superiore di Sanità on the 26th and 27th of November, 2014. The RD-BYOD was attended mainly by RD-Connect partners [95], including rare disease data owners and software engineers with Linked Data expertise. The main focus was to train data owners in making rare disease patient registries and biobanks interoperable, while also identifying tools to be developed.

With support of the BYODs, the rare disease community quickly acknowledged the importance of data interoperability, and later the FAIR principles. In 2017, the International Rare Disease Research Consortium (IRDiRC) declared the FAIR guiding principles as a ‘recognised resource’ to “accelerate the pace of translating discoveries into clinical applications” [96]. Since 2019, the series of annually recurring RD-BYODs has been an intrinsic part of the annual summer school on Rare Disease Registries. Editions of the course have been approved by the International Conference On Rare

[3] For the sake of readability, the “BYOD for rare disease registries and biobanks” is referred to as “RD-BYOD” in this subsection.

Diseases and Orphan Drugs (ICORD) [97].

Over the years, the RD-BYODs evolved to alleviate the steep learning curve of FAIRification. For instance, trainers were instructed to avoid very technical terms that could confuse beginners or participants with different expertise. Additionally, the RD-BYOD has evolved in response to feedback from participants and advancements in FAIR procedures and technologies. For example, training on FAIRification project management for registry managers was added in 2016 and expanded in subsequent editions of the workshop. As a result, from 2017, priority was given to attendees who were involved in or actively planned to establish a rare disease registry, primarily within a European Reference Network (ERN) [28], shifting the focus from a data-focused to a management-focused structure.

The RD-BYOD has also been used to experiment with, get feedback on and disseminate the technical developments that support the rare diseases community. It also informs registry managers about the available tools and standards. Recent RD-BYODs have been adapted to reflect practical aspects of the rare diseases domain, such as including topics to address needs from patient organisations and ERNs. As an example, the EJP RD ontological model for "Common Data Elements", its supporting tool [20], and the EJP RD ontological metadata model for rare disease patient registries, biobanks and catalogues [74] were presented to participants in the latest editions, together with hands-on sessions for demonstration. The experience acquired by RD-BYOD trainers has been embedded in guidance resources, such as a guide for data stewards to make European rare disease patient registries FAIR [21].

3.4 Discussion

The BYODs have benefited attendees and trainers in many ways. For trainers, the workshops have created a collaborative environment where the FAIR community gains new insights from the attendees while helping them deal with their FAIR(ification)-related needs. For example, researchers on FAIR use the open and flexible BYOD environment to test FAIR-related tools with attendees. Similarly, feedback and questions raised during BYODs have supported research on FAIR and FAIRification methods. For instance, research on goal-based FAIRification planning methods [98], assessment of RDF data [99], large-scale implementation of FAIR principles [100] and quality of modelling [101] has benefited from experience from recent BYODs.

Furthermore, lessons learnt from success and shortcomings of BYODs provide guidance on future research paths. To illustrate, the pre- and post-BYOD activities have underscored the iterative nature of FAIRification, where the target resource is initially addressed and then expanded by subsequent FAIRification efforts. For example, in the first FAIRification iteration, a subset of data concepts within a given dataset may be addressed, with subsequent iterations expanding the scope to encompass a larger set of concepts. Other challenges, such as solving the communication gap due to the interdisciplinary nature of FAIR and the diverse expertise of BYOD attendees, highlight the need for further research on such topics. Additionally, the difficulty in reaching consensus during conceptual modelling tasks [102], which are crucial in FAIRification [100], is another obstacle frequently encountered in BYODs.

For attendees, the workshops have aided the participating community by fostering the convergence of standards and tools. In this way, BYODs have become a valuable resource for advancing FAIR data practises. The BYODs' structure has inspired various FAIR training activities and courses, some of which are already offered by universities, other research institutes, or consortia as part of their research data management programmes (e.g. [103, 104, 105]). The Metadata for Machines workshop (M4M) [106] and the Three-Point FAIRification Framework (3PFF) [107] are also examples of a training frameworks that were inspired by the BYODs. Similarly, other FAIRification workflows and frameworks have embedded knowledge acquired by researchers who participated in the BYODs (as trainers or attendees). Examples of these include the generic workflow for the Data FAIRification process [2], the *de novo* FAIRification process of a registry for vascular anomalies [35], the FAIR in action framework for guiding FAIRification [108] and the FAIR Hourglass model for FAIRification and FAIR orchestration [109].

It is also noticeable that BYODs reflect the maturing of the FAIR community. While early BYODs used prototype tools designed to handle specific FAIRification tasks, recent BYODs have introduced more comprehensive tools that can cover a greater part of FAIRification. For instance, while the first BYODs for rare diseases reused generic tools for converting small datasets to linked data, recent ones introduced software systems that can automatically make data FAIR upon collection (e.g. Castor, MOLGENIS). Moreover, recent BYODs were capable of presenting more complex real-world FAIRification cases that lead to new insights and facilitated the retrospective FAIRification of a patient-led registry (e.g. The Duchenne Data Platform [110]).

For future training activities, we recommend combining different types of BYODs to tackle various tasks needed at different stages of FAIRification projects. Practically, a FAIRification project starts by creating a homogeneous basic knowledge about FAIR among all people involved, thus making the commitment and investment efforts clear to the whole FAIRification team. This can be done with management-focused BYODs. After the FAIRification project has been set up and its objectives have been identified, it is necessary to gather sufficient technical expertise, which can be supported by organising data- and software-focused BYODs.

For future BYODs, we plan to explicitly align the contents with the knowledge units mentioned in Appendix E of the FAIRsFAIR Teaching and Training Handbook [111]. Furthermore, we aspire to make our teaching materials themselves FAIR, in order to contribute to overcoming the shortage of FAIR expertise and to continue to keep learning as instructors. We suggest that different communities share training materials and lessons learned, so that BYOD continues to evolve as a whole. In our experience we have also observed that trainers, who are usually FAIR enthusiasts, are willing to support other groups in organising their own BYODs, for example by attending certain BYOD sessions as invited speakers or by giving advice on organising the BYODs.

Finally, we note that BYODs should not be equated with FAIRification projects, as their primary emphasis is on participants rather than output. Nevertheless, BYODs can act as a catalyst for such projects, for example by providing a launch pad for dedicated teams to continue the FAIRification process.

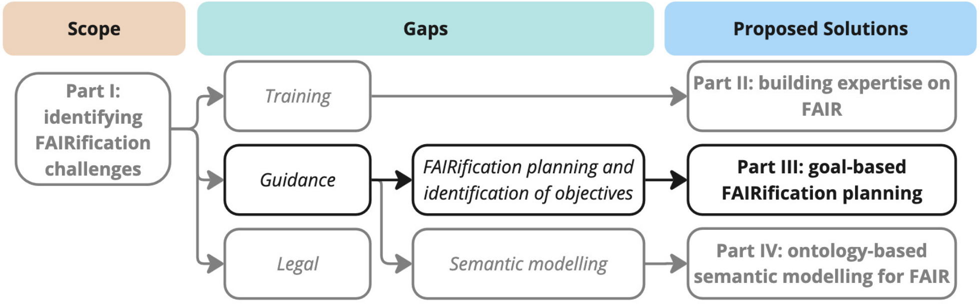
3.5 Conclusion

Initially, BYODs aimed at making data interoperable by using available Linked Data technologies. Since their inception, BYODs have evolved and provided a collaborative space to develop FAIRification tools and more robust technologies. Additionally, BYOD workshops have become an important means of exchanging views and knowledge on FAIRification, and on informing researchers and managers on the benefits of FAIR.

Experience has shown that FAIR implementations are an effective approach to enable multi-source analysis, and the BYODs are a valuable asset in promoting the adoption of the FAIR principles in various domains. We will, therefore, continue to organise BYODs to accelerate the adoption and promotion of the FAIR principles.

Part III

Goal-based FAIRification planning


Scope
Gaps
Proposed Solutions
Part I: identifying FAIRification challenges
Training
Guidance
Legal
FAIRification planning and identification of objectives
Semantic modelling
Part II: building expertise on FAIR
Part III: goal-based FAIRification planning
Part IV: ontology-based semantic modelling for FAIR

Topics covered Part III of this thesis.

In this part, a method to support FAIRification planning through the identification of FAIRification objectives is introduced, addressing the need for guidance in the process identified in Part I. Chapter 4 describes GO-Plan, the goal-oriented method for FAIRification planning. The method builds upon prior experiences with FAIR and FAIRification, as discussed in Parts I and II. It incorporates practices and principles from software engineering and has been validated through two workshops and applied in a real-world scenario. Validation data suggest that end-users find GO-Plan useful and easy to use. The real-world application shows that goal-based diagrams facilitated communication among stakeholders and enabled them to establish achievement criteria for principles that previously lacked precision.

Chapter 4

GO-Plan: A goal-oriented method for FAIRification planning

César Bernabé, Isadora Valle, Tiago Prince Sales, Erik Schultes, Niek van Ulzen, Vítor E. Silva Souza, Annika Jacobsen, Luiz Olavo Bonino da Silva Santos, Barend Mons and Marco Roos

Information Services & Use, (2025).
https://doi.org/10.1177/18758789251324008

Abstract

The FAIR Principles provide guidance on improving the Findability, Accessibility, Interoperability, and Reusability of digital resources. Since the publication of the principles, several workflows have been proposed to support the process of making data FAIR (FAIRification). However, to respect the uniqueness of different communities, both the principles and the available workflows have been deliberately designed to remain agnostic in terms of standards, tools, and implementation choices. Consequently, FAIRification needs to be properly planned, and implementation details must be discussed with stakeholders and aligned with FAIRification objectives. To support this need, this paper describes GO-Plan, a method for identifying and refining FAIRification objectives. Leveraging on best practices from requirements and ontology engineering, the method aims at incrementally elaborating the most obvious aspects of the domain (e.g. the initial set of elements to be collected) into complex and comprehensive objectives. The definition of clear objectives enables stakeholders to communicate effectively and make informed implementation decisions, such as defining achievement criteria for distinct principles and identifying relevant metadata to be collected. GO-Plan has been validated in multiple discussion sessions with experts on FAIR, in an application to a real use case and in two hands-on tutorials with end-users.

4.1 Introduction

The vast amount of data generated every day is only valuable if it can be properly interpreted and reused. However, it is humanly unfeasible to manually merge and make sense of all the information currently available, therefore the support of machines is required. Although machines can automatically analyse and interpret data to efficiently find useful information, they still require time-consuming human support to prepare and merge data [112]. To address this, the FAIR principles have been proposed to guide the transformation and production of resources that are findable, accessible, interoperable and reusable by humans and machines [1]. FAIR resources can be easily managed by machines with minimal human intervention. Nevertheless, the implementation-neutral nature of the FAIR principles makes their realisation a task that requires proper planning and alignment with FAIRification objectives.

The four letters of FAIR are further decomposed into 15 principles [1]. Findability is enforced by using globally unique and persistent identifiers to refer to data and metadata (F1), describing data with rich metadata (F2), explicitly associating metadata with data (F3), and indexing metadata in searchable resources (F4). Accessibility is achieved by using standardised, open communication protocols for data exchange (A1, A1.1) that allow access authorisation procedures (A1.2) while ensuring the longevity of metadata (A2). Interoperability is enhanced by publishing metadata and data in broadly applicable knowledge representation languages (I1), reusing vocabularies that also follow the FAIR principles (I2), and including qualified references to other metadata and data (I3). Finally, reusability is facilitated by describing metadata and data with accurate and relevant attributes (R1), including usage licences (R1.1), detailed provenance (R1.2) and using domain-relevant community standards (R1.3).

The process of making data FAIR (‘FAIRification’) is organised in steps by FAIRification workflows (e.g. [2, 35, 113]). Nonetheless, neither the FAIR principles nor the FAIRification workflows mandate the use of any specific standard, format or software. This is because FAIR and FAIRification have been made agnostic to respect the unique requirements and needs that different communities face when managing and sharing data. Therefore, FAIR can be implemented in different manners and at different levels. However, this flexibility requires careful guidance throughout the FAIRification process to ensure that the implementation decisions (e.g. standards, metadata) align with the FAIRification objectives. In fact, the identification of FAIRification objectives is the initial and crucial step of several FAIRification workflows [114].

The problem of identifying objectives and requirements has been studied by many research communities (e.g. [115, 12]). Among them is requirements engineering, a community dedicated to study the identification, refinement, and management of software requirements [12]. The requirements engineering literature informs that the lack of proper planning and refinement of goals and requirements has a significant impact on the software development process. For instance, Pressman [115] points out that changing requirements after the software product has been delivered can cost up to 60 to 100 times more than changing a requirement during the software planning phase. We hypothesise that the inadequate identification of FAIRification objectives may have a similar impact on planning and executing a FAIRification process. However, there is a scarcity of research on methods specifically focused on supporting FAIRification planning via the identification and refinement of FAIRification objectives. Furthermore, recent studies on the challenges of making rare diseases patients data FAIR have identified that clarifying FAIRification objectives prior to implementation is a key step in FAIRification, as it helps the FAIRification team to make informed decisions that are consistent with their objectives [76].

To address the aforementioned gap, we developed the **G**oal-**O**riented FAIRification **Plan**ning method (GO-Plan) to plan FAIRification through a systematic identification and refinement of FAIRification objectives. The method reflects our understanding that distinct objectives can have different impacts on the planning and execution of FAIRification. Consequently, resources should be made FAIR at a level that aligns with the specific objectives of the FAIRification project. That is, resources should be made "FAIR enough" to fulfil the objectives of the involved collaborators.[1] Thus, the FAIRification planning should not only focus on the selection of suitable technologies or standards, but also on prioritising the effort required to raise the FAIR level of the targeted resources for the realisation of the related objectives. Moreover, as FAIRification is a community-driven, aspirational and incremental process [1], these objectives must encompass the perspectives of collaborators directly participating in the project, but also relevant external stakeholders (i.e. those who will eventually reuse the resource). As such, each effort undertaken to make a resource FAIR (or more FAIR—FAIRer) for one's own objectives will also make that resource FAIRer for others.

This paper is an extension of [98], and it is evolved in five ways. First, we incorporate information on how GO-Plan aligns with the broader FAIRification project. Second, we provide a more comprehensive overview of the types of project collabo-

[1] When referring to *collaborators*, we align with the understanding of the similar term "stakeholder", given by [116] as individuals, groups, or organisations that affect or are affected by a given project.

rators needed at each phase of the method. Third, we report on validation sessions undertaken to assess user perceptions of GO-Plan. Fourth, we present an updated version of the method in this paper. The current version has been adjusted based on the users feedback collected during the validation sessions. Finally, we provide an expanded discussion of related works.

The remainder of this paper is organised as follows. In Section 4.2, we detail the development process of GO-Plan. Following that, in Section 4.3, we describe the method and illustrate it with a fictitious running example. Next, Section 4.4 presents the three validation tasks of GO-Plan. We then discuss related works in Section 4.5. Finally, Section 4.6 discusses the implications of this work for future research.

In this work, we use the spelling "(meta)data" to refer to both data and metadata. The concept of metadata is understood here as "structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource" [117]. Moreover, the words "goal" and "objective" are used in this text as synonyms. Note that the literature on FAIRification workflows usually uses the word "objective", while the requirements engineering literature usually uses the word "goal". Finally, as the scope of FAIR extends beyond data to several types of artefacts (e.g. software, ontologies, workflows), we use the broader term "resource" (e.g. "FAIR resource", "resource made FAIR") to refer to anything that can be made FAIR.

4.2 Development of GO-Plan

Figure 4.1 presents an outline of the process followed in the development of GO-Plan. This process is organised in three main steps, where the third step is further refined into more specific ones.

Identifying the problem The need for a *systematic approach to FAIRification planning* was firstly observed in our collaborations in FAIRification projects. These collaborations included training on FAIR (e.g. [9]), providing guidance to FAIRification (e.g. [118]), and conducting FAIRification within single (e.g. [119]) and across multiple institutions (e.g. [76]). For example, during training sessions aimed at guiding individuals through the steps of FAIRification [9], we observed that participants often encountered difficulties in identifying FAIRification objectives, defining the scope of FAIRification, reaching consensus on the data elements to be collected, and selecting the best implementation solutions based on their FAIRification objectives.

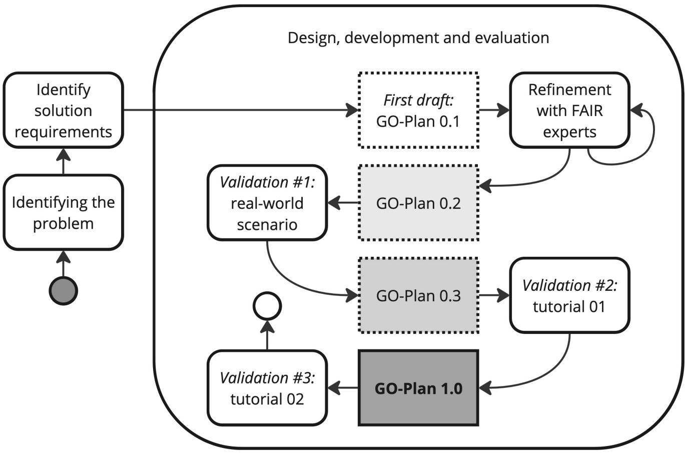

Design, development and evaluation
Identify solution requirements
Identifying the problem
First draft: GO-Plan 0.1
Refinement with FAIR experts
Validation #1: real-world scenario
GO-Plan 0.2
GO-Plan 0.3
Validation #2: tutorial 01
Validation #3: tutorial 02
GO-Plan 1.0

Figure 4.1: Illustration of the steps taken into the development of GO-Plan.

Identifying solution requirements After identifying the problem, we defined base-line requirements for our solution. These requirements can be described as features (Functional Requirements – FRs) and qualities (Non-functional Requirements – NFRs) that GO-Plan must have. In terms of features, we aimed to maintain a sufficient level of versatility so that GO-Plan can be used on different domains (e.g. health care, accounting), at different resources (e.g. data, ontologies, software), and in different types of FAIRification, which include retrospective [2] (when an existing resource is made FAIR), and *de novo* FAIRification [35] (when data is made FAIR automatically upon collection). In terms of desired qualities, we aimed at creating a method that is friendly to newcomers, and therefore it had to be as simple but also as detailed as possible. These requirements are detailed in Table 4.1.

Design, development and evaluation The same experiences that led us to identify the need for GO-Plan also provided tacit knowledge that was embedded in the method. As illustrated in Figure 4.1, the method's first version (GO-Plan 0.1) was further developed throughout several discussions with other experts on FAIR. Subsequently, when a sufficiently stable version of GO-Plan was achieved (GO-Plan 0.2), we proceeded to apply it in a real-world scenario (described in Section 4.4.1). This also prompted for adjustments, resulting in GO-Plan 0.3.

After refining the method following its real-world application, we organised two hands-on tutorials at distinct conferences to evaluate GO-Plan with practitioners (described in Sections 4.4.2 and 4.4.3). During these sessions, we used the Technology

Table 4.1: Requirements aimed during the development of GO-Plan. FR stands for Functional Requirements, which are desired features in the context of this work. NFR stands for Non-functional Requirements, which are desired qualities in the context of this work.

Req. ID	**Requirement Description**
FR-1	GO-Plan can be applied in various domains.
FR-2	GO-Plan can be applied in retrospective and *de novo* FAIRification projects
FR-3	GO-Plan can be applied to different resources.
NFR-1	GO-Plan should be simple enough to be used by individuals with limited expertise in FAIRification projects.
NFR-2	GO-Plan should be detailed enough to provide guidance to individuals with limited expertise in FAIRification projects.

Acceptance Model (TAM) [120, 121] to assess the users' perception of the method in terms of its usefulness and ease of use. The feedback from the first tutorial session led to further adjustments to GO-Plan. As a result, the latest iteration (GO-Plan 1.0), which is presented in this paper, underwent final validation during a subsequent tutorial session.

4.3 The goal-based FAIRification planning method

GO-Plan supports FAIRification planning by systematically defining mature FAIRification objectives through iterative steps. It initially targets the most visible characteristics of the FAIRification project, such as the project domain, scope and available resources. It then leverages these characteristics to address more complex aspects such as relevant data concepts and competency questions. Finally, by following this structured and incremental approach, the method guides collaborators towards the definition of comprehensive objectives that encompass all relevant aspects of FAIRification. These objectives are aligned with implementation decisions, thus resulting in a FAIRification plan.

Figure 4.2 illustrates how GO-Plan should be integrated into the entire FAIRification project. The method should be applied from the moment when the FAIRification project has already been ideated as part of a main project. That is, a project that will use the FAIR resource to fulfil higher-level goals. For instance, when a hospital has decided to create a FAIR patient registry to foster research on rare diseases. At this stage, it is assumed that some aspects, such as the group of people that will be involved in the FAIRification project and the target resources, have already been

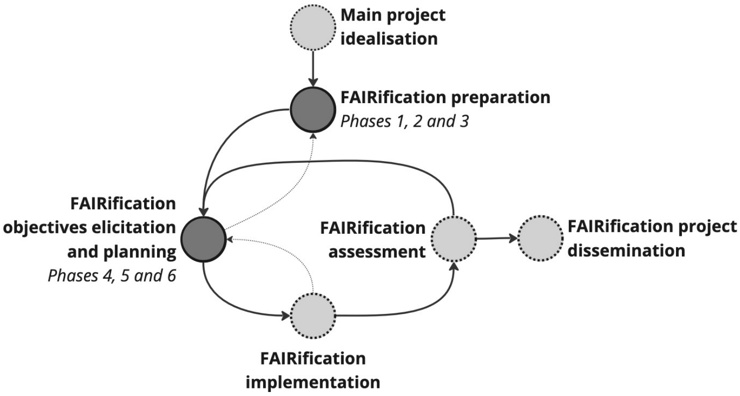

Main project idealisation
FAIRification preparation
Phases 1, 2 and 3
FAIRification objectives elicitation and planning
Phases 4, 5 and 6
FAIRification assessment
FAIRification project dissemination
FAIRification implementation

Figure 4.2: Overview of the placement of GO-Plan in a FAIRification project. Circles with dotted lines and light grey colouring represent activities that are outside the scope of GO-Plan. Arrows with solid lines and dark grey colouring indicate the natural direction of the project, while arrows with dotted lines indicate directions that may be triggered by the need for adjustments or corrections.

initially defined. After the conception of the main project, phases related to the preparation of FAIRification are carried out (phases 1-3, detailed in subsections 4.3.1–4.3.3). These phases consist of collecting information that will support the identification of objectives, as well as preparing the collaborators for the following phases.

After preparation, phases related to FAIRification objectives elicitation and planning take place (phases 4-6, detailed in subsections 4.3.4–4.3.6). These phases consist of elaborating domain concepts, research questions and/or business goals to identify FAIRification objectives that are then formulated into a FAIRification plan. This plan is then used to guide the FAIRification implementation (which is beyond the scope of this work). We recommend the use of FAIRification workflows to support the FAIRification implementation(e.g. [2, 35, 113]).[2]

As illustrated in Figure 4.2, the execution of the FAIRification project is best suited for an agile setting. We suggest that, instead of tackling all objectives concurrently, the team strategically defines a specific set of objectives to be implemented in the current iteration. This approach may involve targeting specific parts of the resource to be made FAIR in each iteration (e.g. focusing on a subset of the whole data, focusing first on metadata). Consequently, the set of objectives will be incrementally elabo-

[2] Note that most FAIRification workflows begin with steps aimed at identifying FAIRification objectives, which may overlap with the aims of GO-Plan. However, these steps are often not detailed or comprehensive enough. Users can substitute these initial steps with GO-Plan for a more structured approach and then proceed with the remaining workflow for FAIRification implementation.

rated upon until all aspects are comprehensively addressed. Moreover, throughout the identification of objectives and the implementation phases, it may become necessary to revisit preceding phases. For example, additional information may need to be gathered during the FAIRification preparation phase to facilitate the further refinement of specific FAIRification objectives. For the sake of simplicity, the remainder of this text delineates the phases and steps sequentially (as one iteration of the agile circular approach).

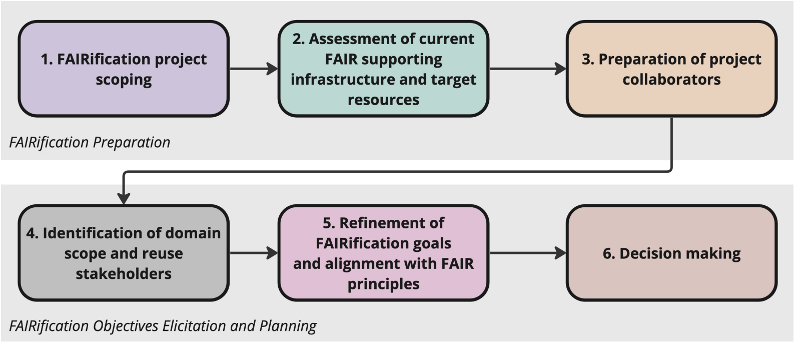

1. FAIRification project scoping
2. Assessment of current FAIR supporting infrastructure and target resources
3. Preparation of project collaborators
FAIRification Preparation
4. Identification of domain scope and reuse stakeholders
5. Refinement of FAIRification goals and alignment with FAIR principles
6. Decision making
FAIRification Objectives Elicitation and Planning

Figure 4.3: The GO-Plan phases are divided into two parts: FAIRification preparation, which consists of the FAIRification project scoping, assessment of current FAIR supporting infrastructure and target resources, and preparation of project collaborators phases; and FAIRification objectives elicitation, which consists of the identification of domain scope and reuse stakeholders, refinement of FAIRification goals and alignment with FAIR principles, and decision-making phases.

As described in Figure 4.3, GO-Plan is organised in six phases, namely (i) FAIRification project scoping, (ii) assessment of current FAIR supporting infrastructure and target resources, (iii) preparation of project collaborators, (iv) identification of domain scope and reuse stakeholders, (v) refinement of FAIRification goals and alignment with FAIR principles, and (vi) decision-making. The phases are refined in several steps and described in the subsections that follow.

A distinction between two categories of collaborators is made throughout the phases of the method: project collaborators and reuse stakeholders. The former refers to those who are directly involved in the FAIRification project and have their own goals and requirements for it (e.g. data custodians, patient representative). The latter refers to those who will eventually reuse the FAIRified resource (e.g. researchers).

It is also important to distinguish the various ways in which collaborators engage in GO-Plan. Firstly, we suggest that all phases are lead by a collaborator with some expertise in objectives elicitation (e.g. requirements engineer, product owner, project leader). Additionally, other roles of collaborators participate in different phases of GO-Plan, as illustrated in Table 4.2. The phases may be supported by advisers, who offer expert guidance without directly having objectives associated with the FAIRification project. Detailed explanations of each collaborator and adviser role are provided in the next subsections.

Table 4.2: Types of collaborators and advisers required in each phase of GO-Plan. While a collaborator would act by providing specific requirements for the FAIRification project, an advisor will support the collaborators by providing expert advice for domain specific or technical questions.

Step	Project collaborator	Adviser
Phase 1	Main project stakeholders (e.g. board members), domain expert, FAIR expert	Maintainers of the resource to be made FAIR, maintainers of the organisation's IT infrastructure
Phase 2	FAIR expert (e.g. FAIR data steward)	Maintainers of the resource to be made FAIR, maintainers of the organisation's IT infrastructure
Phase 3	FAIR experts and domain experts	
Phase 4	FAIR experts and domain experts	Representatives of reuse stakeholders
Phase 5	FAIR experts and domain experts	Goal-modelling experts
Phase 6	FAIR experts and domain experts	Experts on a specific adopted solution

The following subsections describe GO-Plan using a running example of a research organisation that collects data about patients with rare diseases. This organisation has two aims: (i) to make legacy data FAIR (i.e. retrospective FAIRification), and to implement an Electronic Data Capture System (EDC) that already creates FAIR data at the point of collection (i.e. *de novo* FAIRification). In addition to budget and deadline, the most important requirement for this project is the protection of patient privacy through controlled access to the data. The organisation wants to publish non-sensitive data and metadata to foster research on rare diseases.

4.3.1 Phase 1: FAIRification project scoping

As described in Figure 4.4, the method initiates with preparation tasks that entail examining the FAIRification project idealisation documents (e.g. grant proposals, kick-off slides, meeting minutes) and/or holding meetings with project collaborators and advisers (as described in Table 4.2) to identify information that will support subsequent phases. Most of the documentation produced by the method is initially drafted in this phase. The artefacts produced in all phases are summarised and exemplified in Table 4.3.

Table 4.3: List of artefacts produced during the method.

Artefact	Description	Created	Used	Examples
Group of collaborators	List of people actively contributing to the FAIRification project and their roles in the project	1	3, 6	Domain expert (clinician, researcher), FAIR project expert, semantic modeller, developer
Project related goals	Goals that can impact or be affected by the FAIRification, extracted from the goals of a larger project that includes FAIRification	1	4, 6	"Make non-sensitive patient data available Rare Diseases researchers"
Project requirements	Project requirements that will constrain the FAIRification	1	4, 6	Budget, deadline, data privacy constraints, interoperability requirements (e.g. must interoperate with the EJP RD Metadata Model [122])
Resources to be made FAIR	Pre-existing resources that will be made FAIR or modified to generate FAIR data/resources during the FAIRification project	1	2, 4	Legacy data, data collection systems, software, ontology catalogues
FAIR supporting infrastructure	Pre-existing infrastructure that has been allocated to accommodate the FAIR resource	1	2, 6	Current storage servers, access control systems, longevity plans
FAIRification Type	Defines if *post hoc* and/or *de novo* FAIRification must be planned	1	5, 6	-
Domain Description	Describes the domain covered by the FAIRification project	4	4	Description of the set of common data elements for rare diseases registration [20]
Semantic types	Groups of related concepts of similar meaning	4	5	Pain: discomfort, ache, soreness.
Reuse stakeholders	Eventual reusers of the FAIRified resource	4	4, 5	Researchers, data publishers, patient representatives
Competency questions	Questions that must be answered by the FAIRified resource	5	5	see Table 4.4
Metadata concepts	Metadata terms to be collected	6	6	Licence, provenance, related rare diseases codes

In this phase, it is also necessary that the team decides which type(s) of FAIRification should be followed during FAIRification implementation (step 1f): retrospective FAIRification [2], where existing resources are made FAIR, and *de novo* FAIRification [35], where resources are created FAIR (e.g. data made FAIR upon collection).

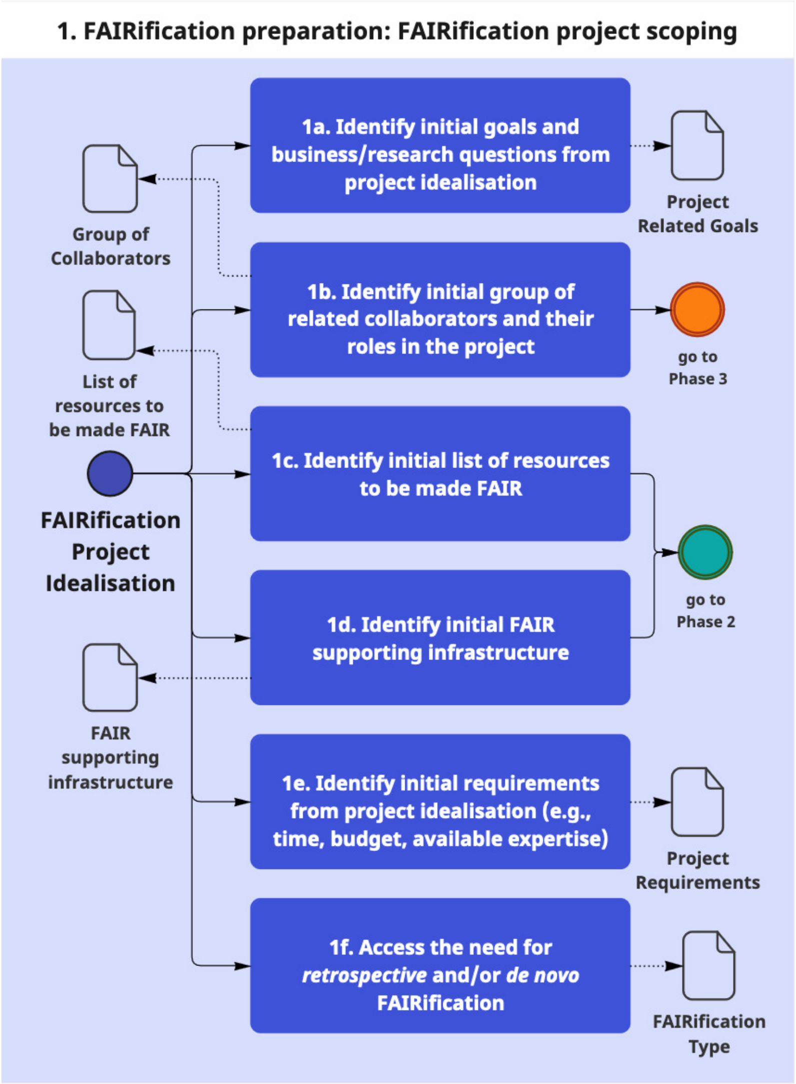

1. FAIRification preparation: FAIRification project scoping
1a. Identify initial goals and business/research questions from project idealisation
Project Related Goals
Group of Collaborators
1b. Identify initial group of related collaborators and their roles in the project
go to Phase 3
List of resources to be made FAIR
1c. Identify initial list of resources to be made FAIR
FAIRification Project Idealisation
go to Phase 2
1d. Identify initial FAIR supporting infrastructure
FAIR supporting infrastructure
1e. Identify initial requirements from project idealisation (e.g., time, budget, available expertise)
Project Requirements
1f. Access the need for *retrospective* and/or *de novo* FAIRification
FAIRification Type

Figure 4.4: Phase 1 of the method. Steps are described inside the phases using the phase number and step letter (e.g. 1a, 1b) as identifiers. The figure follows a BPMN notation [123].

To illustrate, an analysis of the grant application for the rare diseases registry project is conducted to identify relevant stakeholders (step *1b*) and to determine the goals and requirements of the project (steps *1a* and *1e*), as exemplified in Table 4.3. In addition, conducting interviews with project leaders, patient representatives, and researchers can help to identify additional goals and requirements, as well as to identify what resources need to be made FAIR (i.e. legacy patient data and the EDC system) (*1c*). The organisation's information technology (IT) team, together with a FAIR expert, can assist in understanding the existing infrastructure (e.g. storage server for data and metadata, long term longevity plan for metadata) (*1d*) and determining the necessary adaptations required to accommodate the resource to be made FAIR (e.g. changes on the data storage format of the EDC system). Moreover, the data steward of the patient registry can assist in reviewing the data structure and concepts collected by the current registry. Finally, as the project aims both to make legacy data FAIR and to implement an EDC system that already produces FAIR data, the team decides that both retrospective and *de novo* FAIRification must be planned.

4.3.2 Phase 2: Assessment of current FAIR supporting infrastructure and target resources

This phase assesses the resources to be made FAIR and the organisation's currently available FAIR supporting infrastructure. As shown in Figure 4.5, the resources to be made FAIR are assessed (step *2a*) to check if they can be retrieved (e.g. are they in an SQL server hosted locally? In a USB stick at the researcher's home office? Can the current EDC system be modified to generate ontologised data?), understood (e.g. are the headers of CSV files documented? Are the data elements collected by the current EDC system clear enough?) and if there are legal constraints in place (e.g. limited access due to privacy-sensitive data).

Then, the current infrastructure that will accommodate the FAIR resource needs to be assessed (*2c*) to check if it can be used, if it needs to be adapted and/or if additional infrastructure needs to be arranged. The type of infrastructure may vary depending on the type of FAIR resource it is intended to support. For example, to make data FAIR, the infrastructure may include storage servers for data and metadata, and data capturing systems (that might have to be adapted). In the case of privacy-sensitive data, an access control system must be incorporated. Similarly, to make an ontology FAIR, the infrastructure may also involve an ontology repository. In the case of software, examples can include a software code repository and a version control system.

The primary aim of this phase is to ensure that both the resources to be made FAIR and the current infrastructure intended to accommodate the FAIR resource do not pose any obstacles to FAIRification. If any issues are identified in this phase, they must be addressed before continuing to the next phase (steps *2b* and *2d*). If issues found in this phase cannot be solved, then the team must rethink the project idealisation. As shown in Table 4.2, FAIR experts, maintainers of the resource to be made FAIR and maintainers of the organisation's information technology (IT) infrastructure should support this phase.

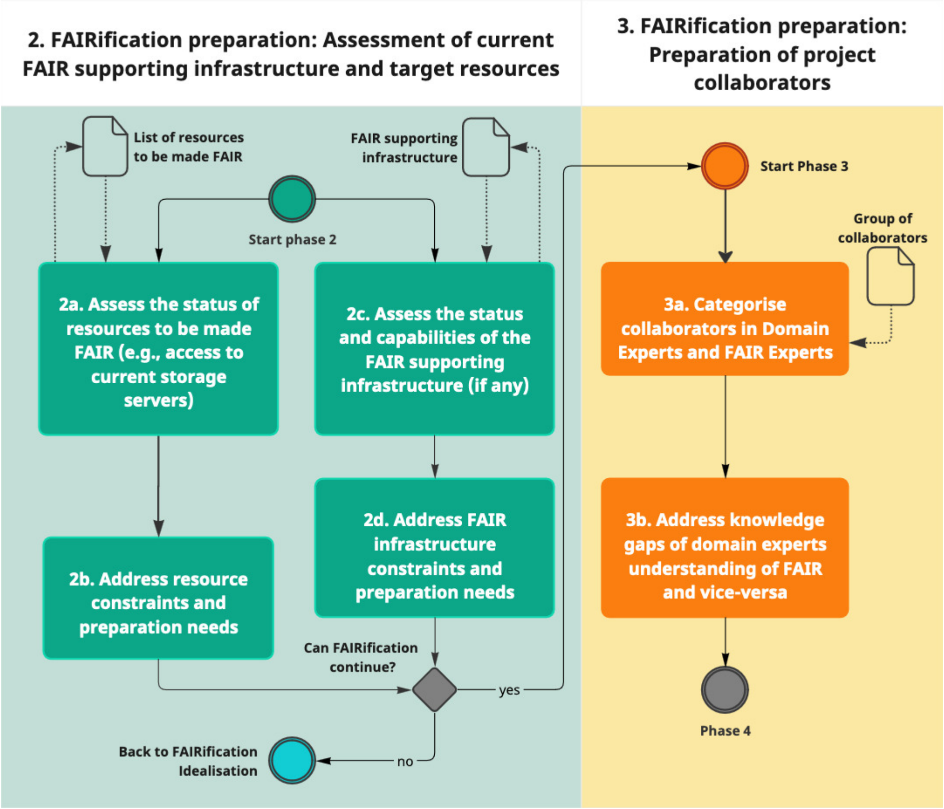

2. FAIRification preparation: Assessment of current FAIR supporting infrastructure and target resources
3. FAIRification preparation: Preparation of project collaborators
List of resources to be made FAIR
FAIR supporting infrastructure
Start phase 2
2a. Assess the status of resources to be made FAIR (e.g., access to current storage servers)
2c. Assess the status and capabilities of the FAIR supporting infrastructure (if any)
2d. Address FAIR infrastructure constraints and preparation needs
2b. Address resource constraints and preparation needs
Can FAIRification continue?
yes
no
Back to FAIRification Idealisation
Start Phase 3
Group of collaborators
3a. Categorise collaborators in Domain Experts and FAIR Experts
3b. Address knowledge gaps of domain experts understanding of FAIR and vice-versa
Phase 4

Figure 4.5: Phases 2 to 3 of the method.

4.3.3 Phase 3: Preparation of project collaborators

The third phase of the method focuses on identifying and preparing the people who will be involved in the FAIRification project. For this, the list of the initial project collaborators is used. The main aim of this task is to bridge the knowledge gap between domain and FAIR experts to prepare them for subsequent phases. The motivation for this comes from the work of Neuhaus & Hastings [102], who suggests techniques to involve stakeholders in the ontology development process. By engaging the project collaborators into each other's domain, we reuse the authors' proposed techniques to establish a more inclusive participatory environment for the discussion of objectives.

In this phase, the group of project collaborators is identified (*3a*) and categorised into FAIR experts and domain experts (*3b*). Then, relevant knowledge gaps between them are assessed to an extent that allows for sufficient understanding of each other's expertise (*3c*). This will create a common "ground language" for stakeholders to communicate their own objectives.

To exemplify, FAIR experts involved in our example project (i.e. rare disease registry FAIRification) could have a question-and-answer session with domain experts about common data elements for rare disease registration [37]. Meanwhile, domain experts get a short lecture on the basics about the FAIR principles and what can be expected and done with FAIR data. We outline that, for the sake of expectation management, it is important to inform domain experts about what is possible with FAIR and what should not be expected as output from a FAIRification project. For instance, while FAIR data may facilitate it, a data visualisation dashboard is a unusual output of FAIRification.

4.3.4 Phase 4: Identification of domain scope and reuse stakeholders

Phase 4 relies on the premise that reuse is the ultimate aim of FAIR, and therefore the FAIRification plan must consider eventual scenarios in which the resource will be used for other purposes. As shown in Figure 4.6, the list of the project goals and the assessment of the resources to be made FAIR are input in this phase to identify and describe the domain scope (*4a*). For instance, rare diseases are the domain of the rare disease registry FAIRification project, while the scope refers to a subset of the domain that considers only the terms of interest for the FAIRification project (e.g. information from patients with rare diseases including treatment procedures may be within the scope, while other medical information unrelated to the rare disease might be out of the scope). Based on our experience, we have observed that using mind maps can facilitate the execution of step *4a*.

It is worth noting that there may be differences between the scope of the FAIRification project and the scope of the resource to be made FAIR. Ideally, these two should be aligned, but this may not always be the case. For instance, the patient registry of our running example may not collect genetic information, whereas this would be important to answer the research question of the FAIRification project. In this case, additional resources (e.g. genetic data) could be added to the FAIRification project. Conversely, information about a patient's health insurance (hypothetically available) in the registry may not be necessary for the FAIRification project, and therefore need not be made FAIR.

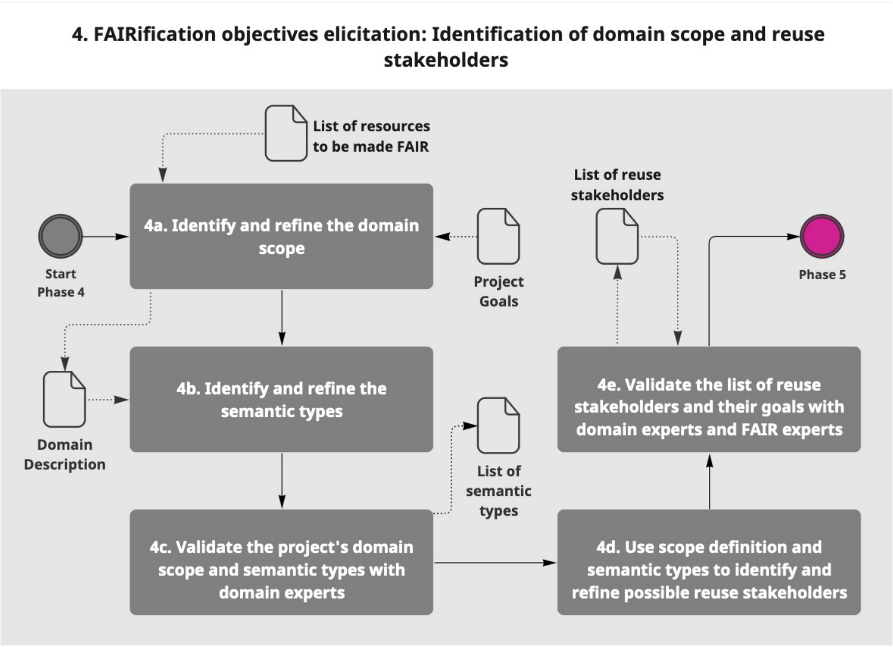

4. FAIRification objectives elicitation: Identification of domain scope and reuse stakeholders
List of resources to be made FAIR
Start Phase 4
4a. Identify and refine the domain scope
Project Goals
List of reuse stakeholders
Phase 5
Domain Description
4b. Identify and refine the semantic types
List of semantic types
4e. Validate the list of reuse stakeholders and their goals with domain experts and FAIR experts
4c. Validate the project's domain scope and semantic types with domain experts
4d. Use scope definition and semantic types to identify and refine possible reuse stakeholders

Figure 4.6: Phase 4 of the method.

Phase 4 also consists of identifying semantic types pertaining to the scope (*4b*). We refer to semantic types as groups of key concepts of similar meaning (e.g. *pain* is a semantic type group that covers similar concepts such as *discomfort*, *ache*, and *soreness*). In our running example, semantic types would include *patient*, *treatment*, *diagnosis* and *genetic information*. These would also be useful in later stages of FAIRification (i.e. conceptual modelling of (meta)data). Next, on step *4c*, the semantic types and their definitions are validated by the group of domain experts. During validation, they may identify additional semantic types to be added to the list.

In step *4d*, the description of the domain and semantic types is used to identify reuse stakeholders. To illustrate, a researcher and a healthcare provider are examples of stakeholders who will reuse patient, diagnosis and treatment data from the rare disease patient registry. Other examples of reuse stakeholders can be patient representatives, clinicians and pharma companies. The list of reuse stakeholders and their goals should be validated with domain experts and also with representatives of the groups of reuse stakeholders themselves (*4e*).

Note that, in step *4d*, it should not be expected a fully comprehensive list of stakeholders, as it would be very difficult to predict all eventual groups of re-users. However, the FAIR project stakeholders should strive for creating a list that considers relevant expected cases. In our real-world experience, we observed that preparing the resource for possible reuse scenarios has a significant impact on the outcome of FAIRification. We also point out that later project extensions to incorporate more reuse cases should be technically feasible given the technical flexibility of FAIR resources.

4.3.5 Phase 5: refinement of FAIRification goals and alignment with FAIR principles

As depicted in Figure 4.7, the fifth phase of the method starts by reusing the list of semantic types defined in the previous phase to identify competency questions (CQs) [124] that should be answered by the FAIR resource (*5a*). In the context of a FAIRification project, a CQ should be a question that would be answered in a significantly easier manner by having the FAIR resource, when compared to before. Additionally, we suggest that CQs elicited in this step should be complex enough to connect and explore the relationship between different semantic types. Table 4.4 shows some examples of CQs that can be defined for the semantic types exemplified in Section 4.3.4. In step *5b*, the CQs are assigned to related stakeholders (i.e. reuse stakeholders and relevant project stakeholders) and further refined as objectives (*5c*). These objectives can be identified by asking *why* a certain CQ needs to be answered and *how* it can be answered. Some objectives are also exemplified in Table 4.4. For the objectives related to the reuse stakeholders, it is recommended that they are elicited and validate with representatives of such groups.

Table 4.4: Examples of competency questions (CQs), related stakeholders and their objectives.

CQ	Stakeholder	Objectives
What is the age range and gender distribution of patients with a particular rare disease in Europe?	Patient Representative	“Public awareness of Rare Diseases is improved”
	Health Care Provider	“Patient management is improved”
What previous diagnoses and treatments have been tried for patients with a particular rare disease?	Researcher	“Cohorts for clinical trials are identified”, “Disease progression is predicted”

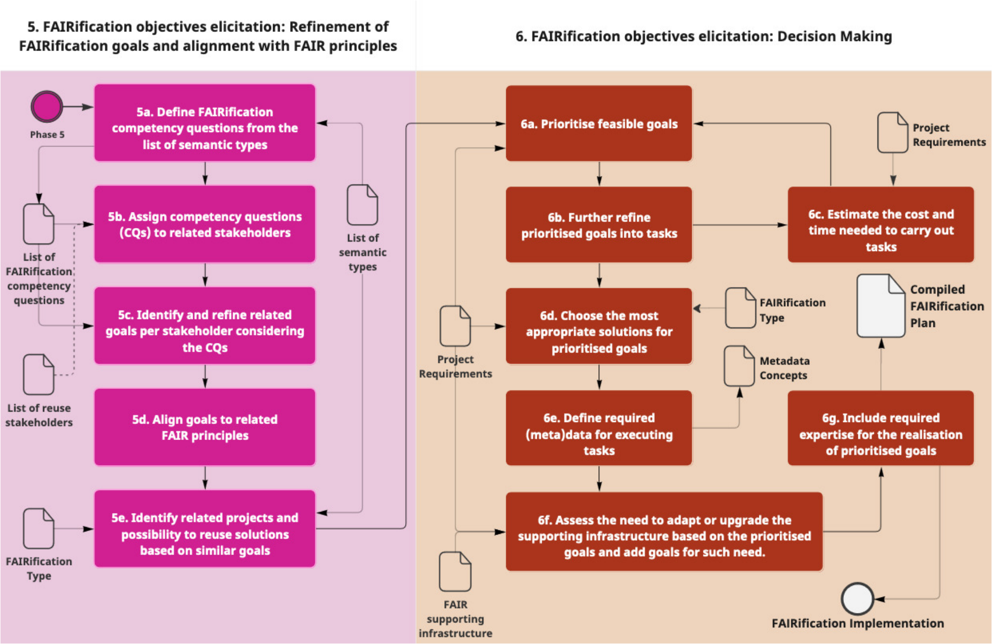

5. FAIRification objectives elicitation: Refinement of FAIRification goals and alignment with FAIR principles
6. FAIRification objectives elicitation: Decision Making
Phase 5
5a. Define FAIRification competency questions from the list of semantic types
5b. Assign competency questions (CQs) to related stakeholders
5c. Identify and refine related goals per stakeholder considering the CQs
5d. Align goals to related FAIR principles
5e. Identify related projects and possibility to reuse solutions based on similar goals
List of FAIRification competency questions
List of reuse stakeholders
FAIRification Type
List of semantic types
6a. Prioritise feasible goals
6b. Further refine prioritised goals into tasks
6c. Estimate the cost and time needed to carry out tasks
6d. Choose the most appropriate solutions for prioritised goals
6e. Define required (meta)data for executing tasks
6f. Assess the need to adapt or upgrade the supporting infrastructure based on the prioritised goals and add goals for such need.
6g. Include required expertise for the realisation of prioritised goals
Project Requirements
Project Requirements
FAIRification Type
Metadata Concepts
Compiled FAIRification Plan
FAIR supporting infrastructure
FAIRification Implementation

Figure 4.7: Phases 5 and 6. The final set of compiled FAIRification objectives is depicted as the white coloured artefact.

Subsequently, the objectives identified from the CQs are aligned with related FAIR principles (*5d*). For this step, it should be identified which and how a principle will support achieving a specific objective. For instance, the objective "public awareness of rare diseases is improved" (Figure 4.8), which is further refined until it can be realised by the task "collect and publish demographic statistics", may be supported by F2 (rich metadata to make the patient registry findable) and R1.1 (data licence to allow reuse of the data for demographic statistics). Meanwhile, other principles (e.g. F1) may not be prioritised for this specific objective.

To facilitate the management of objectives, we suggest the use of goal-modelling techniques such as *iStar* [16], which helps to capture the stakeholders intentions and their relationships in a structured way. Models created with *iStar* include concepts such as *actors*, *goals*, *qualities*, *tasks*, *resources*, and relationships such as *decomposition*, *contribution* and *dependency* links. In the context of a FAIRification project, *actors* represent the collaborators, required expertise and the reuse stakeholders. *Goals* and *tasks* describe the FAIRification objectives and the FAIRification implementation tasks, respectively. *Qualities* describe external goals from the main project. *Resources*

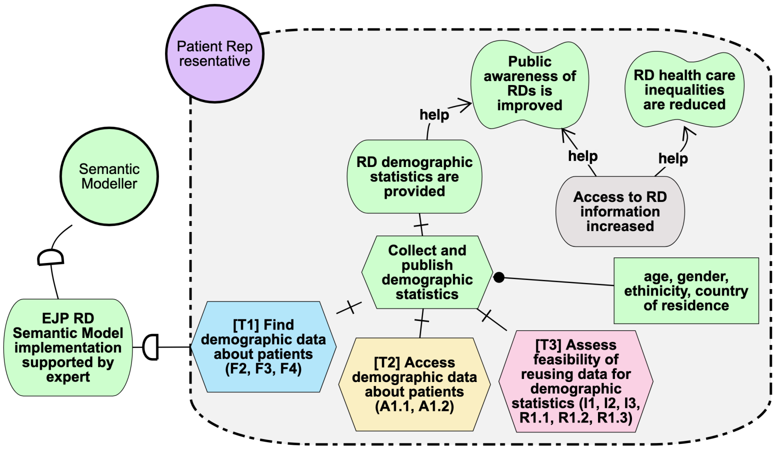

Patient Rep resentative
Semantic Modeller
EJP RD Semantic Model implementation supported by expert
Public awareness of RDs is improved
RD health care inequalities are reduced
help
help
help
RD demographic statistics are provided
Access to RD information increased
Collect and publish demographic statistics
age, gender, ethinicity, country of residence
[T1] Find demographic data about patients (F2, F3, F4)
[T2] Access demographic data about patients (A1.1, A1.2)
[T3] Assess feasibility of reusing data for demographic statistics (I1, I2, I3, R1.1, R1.2, R1.3)

Figure 4.8: Excerpt of the objectives model of a patient representative using *iStar*. Rounded shapes represent *goals* (objectives), cloud-like shapes represent *qualities*, hexagons represent *tasks* and *rectangles* represent resources reused by tasks. Elements in dark grey are not prioritised. The dashed line defines the boundaries of the stakeholder intentions.

can be used to describe the resources to be made FAIR, the solutions to be reused in the FAIRification implementation tasks and the metadata concepts to be collected. *Decomposition* can be used to model how high-level FAIRification objectives are further decomposed into more specific ones. Finally, *Contribution* can be used to describe the contributions among the FAIRification objectives themselves and from these to the main project's goals. *Decompositions* can represent expectations among different contributors. The reader is referred to Dalpiaz *et al.* [16] for further information on *iStar*.

The final step of this phase consists of using the list of semantic types to identify related FAIRification projects (*5e*) through, for instance, the use of FAIR Implementation Profiles (FIPs) [61] or standards catalogues such as FAIRSharing [125]. FIPs are specifications of implementation solutions for realising the FAIR principles in a specific context or domain, and their use is intended to foster convergence on FAIR implementation decisions [61]. In the context of GO-Plan, related projects can support collecting implementation solutions that can be reused in the FAIRification project. The EJP RD project [122] is such a project to our running example.

4.3.6 Phase 6: Decision making

The sixth and last phase of the method starts by prioritising feasible objectives (*6a*) given the project requirements (e.g. data privacy) and constraints (e.g. budget, deadline, available expertise). GO-Plan suggests that this prioritisation is done by (i) refining a set of goals into high-level implementation tasks (*6b*), and then (ii) estimating the cost and time for executing such tasks (*6c*). Thus, prioritisation is done by comparing the cost and time parameters of each task with the project requirements. Additionally, this step must include removing objectives that are not feasible, may not be supported by FAIR principles or are not related to FAIRification.

Next, the most appropriate solutions for prioritised objectives are identified and selected considering the project goals and requirements, and the limitations of available supporting infrastructure and expertise (*6d*). This step can be supported by reusing solutions from the similar projects identified in step *5e* and by consulting experts. Then, the necessary (meta)data for achieving the identified tasks are listed (*6e*) and described as resources in the goal diagrams, as shown in Figure 4.8. The next step (*6f*) is to consider whether the organisation's current support infrastructure needs to be adapted or upgraded in the light of the prioritised goals. If so, objectives to address this need should be added to the list of objectives.

Subsequently, the required expertise for the implementation of the selected solutions (*6g*) is defined. To illustrate, the reuse of the EJP RD Metadata Model is a possible implementation choice for the objectives depicted in Figure 4.8 (in the context of F2 – "Find demographic data about patients") given the project requirements, and a semantic modelling expert would be a required expertise to support reusing this solution.

At this point, the goal diagram should contain enough information to inform and guide FAIRification. The FAIRification objectives, tasks and chosen implementation solutions can now be seen as actions to be taken towards realising FAIRification, that is, the FAIRification plan. It is upon the experts conducting the FAIR project to define implementation cycles and evaluation cases.

As indicated in Table 4.3, various documentation artefacts (e.g. list of collaborators and resources, domain description) are generated throughout the execution of GO-Plan. The design and organisation of these artefacts should be tailored to meet the specific requirements of the FAIRification team. To facilitate this, a comprehensive document containing *suggested* templates for all information produced via GO-Plan has been compiled and is accessible (under CC BY 4.0 licence) at doi.org/10.5281/zenodo.14780348. An illustrative document containing the running example used in this paper is also available at the same link.

4.4 Validation

As described in Section 4.2, GO-Plan has been evaluated from different perspectives. Firstly, it was applied to a real-world scenario and adapted according to the feedback received (described in Subsection 4.4.1). Subsequently, we validated the method in two tutorial sessions (described in Subsections 4.4.2 and 4.4.3). During these tutorials, participants with different levels of FAIR expertise experimented with GO-Plan in a mock case scenario. The method was adjusted after the first tutorial and the current version was tested during the second tutorial.

4.4.1 Application in a real-world scenario

In the validation in a real-world scenario, GO-Plan was used to improve the FAIR level of the OntoUML/UFO catalogue [126, 127], which contains a growing set of conceptual models defined either using the OntoUML modelling language [17] or by extending the Unified Foundational Ontology (UFO) [18]. The catalogue was initially built using *ad hoc* FAIRification,[3] as reported in [126], and later had its FAIR aspects reviewed using GO-Plan, as detailed in [127].

Methods and materials

We structured the validation in a real-world scenario into four distinct steps. Initially, we instructed all collaborators of the FAIRification project in the use of GO-Plan (step 1). These collaborators consisted of individuals involved in the development of the OntoUML/UFO catalogue (i.e. ontologists, developers and FAIR experts). Subsequently, the collaborators autonomously applied the method to their case and constructed a FAIRification plan (step 2). Thirdly, after completing the FAIRification plan, another session was conducted to gather feedback on the collaborators' perception of the method (step 3). In this feedback session, we focused on understanding the differences of using GO-Plan when comparing it with *ad hoc* FAIRification. Fourthly, the feedback gathered during this session guided refinements to GO-Plan (step 4).

[3] Ad hoc FAIRification refers to the process of making resources FAIR on an as-needed basis, rather than through a standardised, systematic approach.

Instructing collaborators in the use of GO-Plan was straightforward, as all them had considerable knowledge in goal modelling, ontology engineering and basic to expert knowledge in FAIRification. The instructions were given in a lecture format, where each step of GO-Plan was explained using the same running example described in this paper (i.e. patient registry).

Results and discussion

We observed that, when applying GO-Plan to their own case, the team could easily navigate through the method's phases 1 to 3 because the FAIRification project of the OntoUML catalogue had already been designed and previously executed (i.e. [126]). Consequently, the team already had a base knowledge of the project's scope, infrastructure, associated collaborators, and so forth. In this case, GO-Plan was used to review and refine these aspects. Additionally, the team did not identify any issues in phase 2 regarding the supporting infrastructure and resources to be made FAIR. Moreover, phase 3 was not necessary as all involved stakeholders already had sufficient knowledge of FAIR and the domain (i.e. OntoUML and UFO).

In GO-Plan's phase 4, the team delimited the scope of the FAIRification project using the OntoUML metamodel, which was also used to define most semantic types of interest (e.g. class, relation, design pattern). In phase 5, the team decided that only the catalogue manager, as a project stakeholder, would be included in the set of goal diagrams alongside the reuse stakeholders. This decision was based on the fact that the catalogue manager's objectives were deemed to be the most pertinent to the project and appeared to encompass the additional objectives of the entire team. For reuse stakeholders, the team identified the ones described in Table 4.5, and defined their goals based on the catalogue manager's objectives. For instance, "Domain model become a community reference" is a goal related to the catalogue manager's objective "model reuse maximised".

Table 4.5: Reuse stakeholders identified in the OntoUML catalogue FAIRification project and the objectives identified for each of them.

Reuse Stakeholder	Initial objectives
Newcomer	"Proficiency in OntoUML increased"
Modeller	"Have domain of interest defined by a model", "Domain model become a community reference", "Domain model reuse maximised"
Tool Developer	"Algorithm developed", "Have algorithm evaluation reproducible"
Researcher	"OntoUML language improved", "OntoUML be used appropriately'

Finally, on phase 6, the team aligned the objectives to related FAIR principles (e.g. Figure 4.9) and defined implementation solutions for each principle in the context of the referred objectives. Although the solutions were not reused from similar projects, the team consulted with experts to identify the most appropriate ones (e.g. using DCAT [128] for defining metadata). A complete description of all objectives pertaining to each stakeholder is available in the project's documentation.[4]

During the validation of the resulting FAIRification plan, the collaborators observed that certain objectives within the resulting model needed to be reassessed for feasibility. This prioritisation was done by the team by analysing the aspects of the FAIRification tasks. For example, the team analysed the time, cost and expertise required to complete the tasks, which helped them to prioritise based on the project constraints.

The feedback session was conducted with representatives of the collaborators of the FAIRification project and the developers of GO-Plan. In this session, we agreed on the need for adjustments on (i) clarifying the description of phases and steps (e.g. *2a* and *2c* were renamed), (ii) reordering the sequence of steps (e.g. *1d* was actually a step before *2c*), and (iii) addressing missing steps (e.g. adding steps *1e*, *1f*, *4e*, *6a* and *6c*).

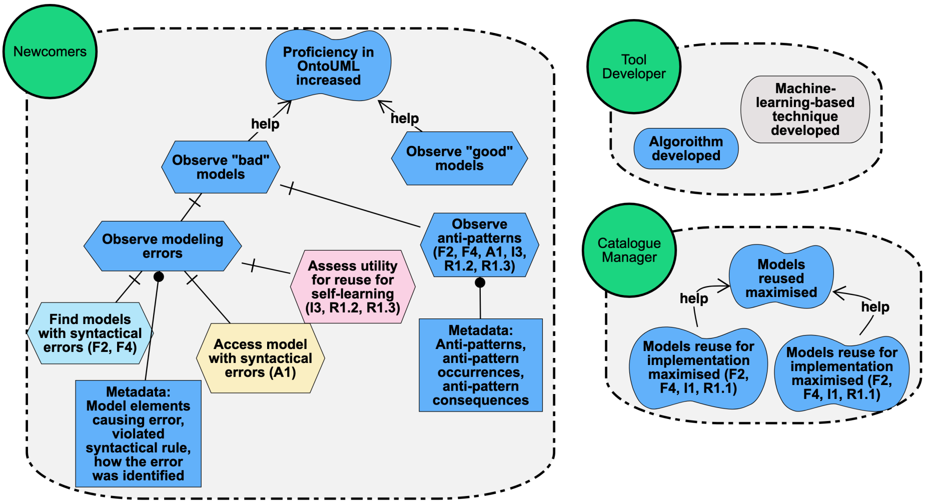

Newcomers
Proficiency in OntoUML increased
help
help
Observe "bad" models
Observe "good" models
Observe modeling errors
Observe anti-patterns (F2, F4, A1, I3, R1.2, R1.3)
Assess utility for reuse for self-learning (I3, R1.2, R1.3)
Find models with syntactical errors (F2, F4)
Access model with syntactical errors (A1)
Metadata: Anti-patterns, anti-pattern occurrences, anti-pattern consequences
Metadata: Model elements causing error, violated syntactical rule, how the error was identified
Tool Developer
Machine-learning-based technique developed
Algoroithm developed
Catalogue Manager
Models reused maximised
help
help
Models reuse for implementation maximised (F2, F4, I1, R1.1)
Models reuse for implementation maximised (F2, F4, I1, R1.1)

Figure 4.9: An excerpt of the objective model diagram defined for the Newcomer and Tool Developer reuse stakeholders.

[4] https://purl.org/ontouml-models

When applying the method to the real-world use case [127] and comparing it to *ad hoc* FAIRification [126], it was observed a significant influence of defining reuse stakeholders in the results of FAIRification. This improvement was particularly noticeable when the team had to identify which (meta)data concepts should be collected and published, as well as considerations regarding licensing and provenance. We attribute this impact to the fundamental emphasis of FAIR on facilitating reusability and assert that optimising the resource for reuse cases is key to effective FAIRification.

The team reported that using goal-based diagrams has facilitated the communication among them. Additionally, they reported that our approach led to more informed and clearer decision-making and evaluation of the FAIRness of the catalogue. The stakeholders were able to prioritise solutions based on a comprehensive understanding of the relationship between objectives and the FAIR principles. To illustrate, the use of our method resulted in a re-definition of metadata concepts to be collected, a re-prioritisation of the principles (e.g. more attention was given to R1), and the inclusion of FAIR supporting infrastructure such as the FDP.

Finally, the collaborators mentioned that the resulting objectives helped stakeholders in establishing achievement criteria for principles that lacked sufficient precision. For instance, the team was able to define that the metadata set satisfied the "data are described with rich metadata" (F2) principle by ensuring that it supported all prioritised goals from the reuse stakeholders.

Threats to validity

Wohlin *et al.* [129] classify four common types of threats to validity found in software engineering experiments. As GO-Plan draws on methods from this field, we adopt these categories to assess the threats to validity inherent to validating the method.

Internal Validity: This aspect pertains to the degree to which a study effectively establishes a causal relationship between observation and results. This includes the risk of making incorrect observations during the application of the method and misinterpreting the feedback received from participants. To mitigate this risk, we conducted an additional session with participants after incorporating adjustments on GO-Plan. This session was intended to check whether our observations were correct, and that the adjustments addressed the team's feedback.

External Validity: External validity refers to the extent to which a study's findings can be generalised to real-world applications. We understand that this validation could have been biased as we opted to engage a group of experts as our primary audience. Ideally, GO-Plan should have been tested with people with varied levels of experience in FAIRification, as put by NFR-1. Still, initially testing the method with an expert audience was deemed important at an early development stage because we wanted to gather more specialised feedback on our methodology. We also acknowledge that GO-Plan should have been validated in different domains to guarantee its generalisability.

Construct Validity: Construct validity assesses the extent to which a study accurately measures the constructs or concepts it is intended to measure. In this perspective, potential biases include the intentional omission of phase 3, which was not actively experienced by the experiment participants. Moreover, an additional concern regards the fact that GO-Plan was compared against an *ad hoc* FAIRification plan. This comparison may be seen as unfair due to the principle that using any method is in most cases better than using no method [130].

Conclusion Validity: Conclusion validity pertains to the accuracy of inferences drawn from the data. This includes the risk of drawing conclusions about the method's validity without using a formal approach for data collection and analysis. Although our conclusions were supported by a partially structured approach (i.e. observation, synthesis, discussion), we did not employ any established methods from the literature to perform this task.

In essence, the threats to validity identified in this validation process could be addressed by considering five key measures: (i) involving users with different levels of expertise, (ii) extending the validation to multiple domains, (iii) avoiding direct comparison with *ad hoc* FAIRification planning methods, (iv) comprehensively validating the method without skipping any phase or step, and (v) using a formal method for data collection and analysis. These considerations were addressed in the design of a second validation strategy, which is described in detail in the following subsection.

4.4.2 Validation in the first tutorial

The second validation activity involved assessing GO-Plan via the "insights in FAIRification planning" tutorial, which was held at the 27th edition of the Enterprise Design, Operations, and Computing (EDOC 2023) conference [131]. Here, the feedback from participants during their use of GO-Plan also prompted for improvements to the method.

Methods and Materials

The tutorial session was initiated with a lecture-style presentation introducing FAIR and FAIRification. This presentation was tailored for those with limited prior knowledge on the subject. Subsequently, we introduced the tutorial mock case, which was followed by a step-by-step explanation of GO-Plan. The explanation of the method was done in incremental parts combined with hands on. Then, after completion of all hands-on parts, participants were invited to complete a questionnaire designed to assess their perceptions of the method. The slides, documentation templates, and all other resources utilised during the tutorial are accessible on the tutorial's website.[5]

Introductory presentations on FAIR and FAIRification The presentation on FAIR and FAIRification was intended to level the participants' understanding up to a minimal ground knowledge on the topic, which consequently also fulfilled GO-Plan's step 3b. The presentations consisted of an introduction on the benefits FAIR, followed by an explanation of the FAIR principles and examples of technologies and artefacts that can be used to achieve FAIR (e.g. ontologies, data formats, the FDP). Furthermore, an overview of FAIRification workflows was provided with an illustrative example based on the generic FAIRification workflow. These lectures were made interactive through the use of voting, experience sharing and Questions and Answers (Q&A) sessions.

The mock case In designing the mock case for the tutorial, we sought to create a scenario that was relatively straightforward, yet one that would resonate with a broad audience. The mock case comprised a FAIRification project for *K-Woef*, a fictitious dog welfare organisation with the objective of increasing dog adoption rates in the United States. *K-Woef*'s aim was twofold: (i) improve data sharing between shelters for more efficient matching of dogs and adopters, and (ii) make parts of the data publicly accessible to influence public policies in favour of animal protection. To replicate potential challenges concerning the FAIR infrastructure and targeted resources (GO-Plan's phase 2), each group received a deck of cards containing randomised situations reflecting possible project obstacles. For instance, one card prompted: "You need to deploy a FAIR Data Point because your organisation does not have one." Some groups received cards requiring them to plan a *de novo* FAIRification implementation, while others received cards instructing them to plan retrospective FAIRification. These cards were carefully designed to ensure uniform difficulty levels across all groups.

Presentations on GO-Plan and hands-on sessions In order to prevent participants from becoming overwhelmed by an excess of information, we combined the

[5] `https://fairification-planning.github.io/`

step-by-step explanations on GO-Plan with hands-on sessions. Firstly, we introduced phases 1 to 3 of the method, and then participants implemented these phases in the mock case. Subsequently, we explained the steps of phase 4 and provided another session for its application to the mock case. Finally, we discussed the last two phases and let participants apply them to the mock case. Each hands-on session lasted 40 minutes. All presentations about the method's phases were followed by Q&A sessions. For the hands-on, participants were organised in groups with similar number of members. Within these groups, members collaborated to apply the method to the mock case scenario. Participants were also provided with a template document (optional use) for describing the artefacts produced during the application of GO-Plan (e.g. list of stakeholders, list of semantic types).

The questionnaire Following the conclusion of the third hands-on session, each group had formulated a FAIRification plan. Subsequently, participants were invited to complete a questionnaire created to gather their perceptions of GO-Plan. For this, we used the Technology Acceptance Model (TAM) [120, 121] as a framework for designing the questionnaire, collecting and analysing the results. TAM has been widely applied and tested in various contexts and domains and is one of the most influential and cited models in the field of information systems and technology acceptance [121].

As illustrated by Figure 4.10, TAM proposes that two factors determine a user's intention to use a technology: perceived usefulness (PU) and perceived ease of use (PEOU). PU is the degree to which a user believes that using a technology will enhance their performance, while PEOU is the degree to which a user believes that using a technology will be free of effort. Additionally, TAM suggests that external variables (EV), such as the system or method features, can affect the user's perceptions of PU and PEOU. Moreover, PEOU is anticipated to influence PU, as achieving higher efficiency in completing tasks with the same effort, facilitated by improvements in a system's usability (reflected in higher PEOU), can directly impact the users' perception of usefulness (PU). The PEOU and PU are expected to impact the Attitude Towards Using (ATU), which is defined as an individual's positive or negative sentiment about using a particular system or method. Lastly, ATU influences the Behavioural Intention to Use (BIU), suggesting that individuals' intentions to engage in certain behaviours are closely linked to their attitudes, which result from their positive perceptions of a system or method. Finally, the PU-BIU relationship represents the direct effect of people's intentions toward using a new method based largely on how it will improve their performance on a certain task.

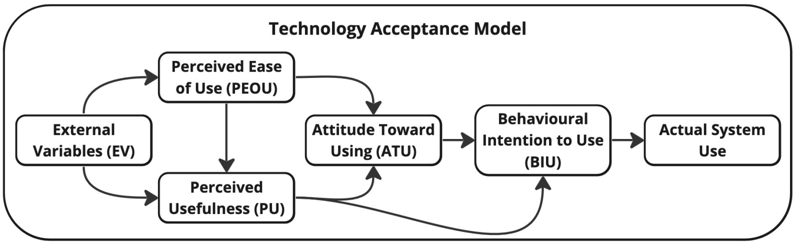

Technology Acceptance Model
Perceived Ease of Use (PEOU)
External Variables (EV)
Attitude Toward Using (ATU)
Behavioural Intention to Use (BIU)
Actual System Use
Perceived Usefulness (PU)

Figure 4.10: Illustration of the Technology Acceptance Model, designed based on the original figure from [120].

As depicted in Figure 4.11, we adapted and simplified TAM to evaluate GO-Plan by incorporating the expected benefits associated with goal-modelling as external variables (EV). These benefits are defined by the work of Horkoff and Yu [132] and described in Table 4.6. Additionally, we excluded the PEOU-PU and PU-BI relationships in our instance of TAM. This decision was deliberate as both are tied to users' perceptions when comparing the method with their own practices, specifically with previous FAIRification projects. Including these relationships could have introduced bias into our results because we could not control the participants' experience levels prior to conducting the tutorial (e.g. ensuring a balanced number of participants with similar experience levels). Due to the nature of conference-based tutorials, we did not have access to this information before receiving the participants on the day of the tutorial.

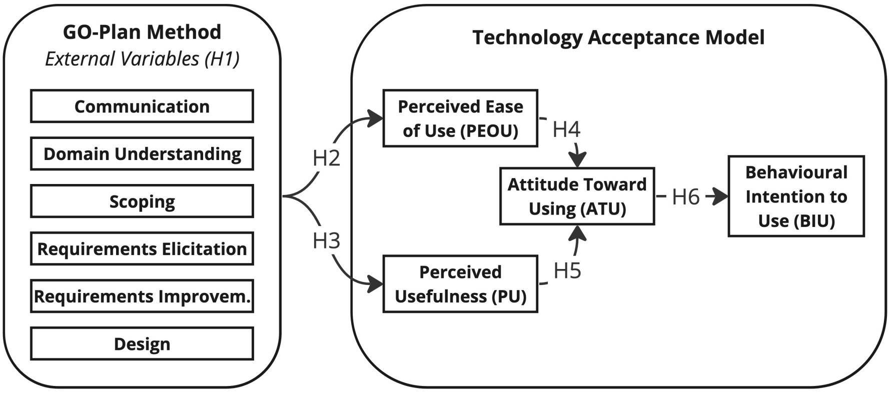

GO-Plan Method
External Variables (H1)
Communication
Domain Understanding
Scoping
Requirements Elicitation
Requirements Improvem.
Design
H2
H3
Technology Acceptance Model
Perceived Ease of Use (PEOU)
H4
Attitude Toward Using (ATU)
H6
Behavioural Intention to Use (BIU)
H5
Perceived Usefulness (PU)

Figure 4.11: The Technology Acceptance Model adapted with goal-modelling characteristics [132] as external variables.

Following the guidelines of TAM, we derived six hypotheses (Table 4.7) from our adapted model. The initial hypothesis (*H1*) aims to verify whether the benefits of

Table 4.6: List of benefits associated with using goal-oriented approaches, as described by Horkoff and Yu [132].

Benefit	**Description**
Domain understanding	Goal-oriented approaches are anticipated to enhance the understanding of the domain being modelled, especially for individuals who are not experts in that particular domain.
Communication	Goal models function as effective tools for communicating and validating concepts within the domain being modelled, also serving as a means of conveying design alternatives.
Scoping	Goal methods support defining the boundaries of the project's scope and delineating the areas of interest of involved collaborators.
Requirements elicitation	Goal modelling contributes to the discovery of requirements, enhancing the comprehensiveness of models and prompting for further elicitation of objectives and tasks.
Requirements improvement	Goal-based procedures help users identify deficiencies within objectives, such as conflicts between them and lack of refinement.
Design	Goal-oriented methods are instrumental in exploring alternative implementation solutions and providing justification for decisions made during the design process.

using goal-modelling translate to GO-Plan. Consequently, H1 is refined into six sub-hypotheses, each addressing a specific benefit mentioned in Table 4.6. Subsequently, we check if the method is perceived as easy to use (*H2*) and useful (*H3*). Further, we test if users' attitude towards using GO-Plan correlates with PEOU (*H4*) and PU (*H5*). Lastly, we evaluate the users' behavioural intention to use GO-Plan in relation to their attitude towards its use (*H6*).

To test our hypotheses, we elaborated 20 close-ended questions to which responses were collected using a 7-point Likert scale (i.e. 1- strongly disagree to 7- strongly agree). Additionally, we included another 8 open-ended questions primarily focused on understanding the participants' background and other perceptions of the method. Examples of questions from the questionnaire are outlined in Table 4.8. The complete list of questions is found in the supplementary material.[6]

[6] `doi.org/10.5281/zenodo.14780348`

Table 4.7: Set of hypotheses derived from adapting TAM for the evaluation of GO-Plan.

Hypothesis	Description
H1	Goal-modelling benefits are perceived by those using GO-Plan.
H1.1	Those using GO-Plan notice that they can better communicate about the FAIRification domain.
H1.2	Those using GO-Plan notice that they can better understand the FAIRification domain.
H1.3	Those using GO-Plan notice that they can better define the FAIRification scope.
H1.4	Those using GO-Plan notice that they can better perform a FAIRification requirements elicitation.
H1.5	Those using GO-Plan notice that they can improve the requirements elicited for FAIRification.
H1.6	Those using GO-Plan notice that they can better design a FAIRification plan.
H2	GO-Plan is perceived as easy to use by its users.
H3	GO-Plan is perceived as useful by its users.
H4	The user's perception of GO-Plan's ease of use positively affects their attitude towards using the method.
H5	The user's perception of GO-Plan's usefulness positively affects their attitude towards using the method.
H6	The user's positive attitude towards using GO-Plan positively impacts their behavioural intention to use it.

Results and discussion

We collected 6 responses in the tutorial. These were from participants with no or little previous experience in FAIRification projects (i.e. 2 had been involved in one FAIRification project before, while 4 were new to the topic). Four participants mentioned to have previous expertise in ontologies and conceptual modelling.

Responses to close-ended questions Table 4.9 describes a simplified statistical analysis (given the small sample size of 6 responses) of the answers to close-ended questions. To calculate the mean for each hypothesis, we first combined related questions to form our theoretical constructs used in the hypothesis testing. We then report the sample mean for each theoretical construct. For example, if Participant-A answered

Table 4.8: Example of questions asked to participants of the tutorial experiment and related hypotheses for each question. The results of these questions are used to validate the related hypotheses.

Question	**Related Hypoth.**
Learning about dog adoption helped me to communicate with other team members about this topic.	H1.1
Learning about FAIR and FAIRification helped me to communicate with other team members about this topic.	H1.1
I have no difficulty understanding GO-Plan's tasks.	H2
I am willing to use GO-Plan in my FAIRification projects in the future because I find the method easy to use.	H4
If I participate in FAIRification projects in the future, I will use GO-Plan in these projects or I will suggest it to other people.	H6
How confident are you in realising FAIRification?	–

'5' to Q1 and '3' to Q2, which are both questions associated with *H1.1*, then we calculated '4' (i.e. (5+3)/2) as Participant-A's score for *H1.1.* If participant B answered '3' to Q1 and '1' to Q2, we calculated '2' (i.e. (3+1)/2) as the score for *H1.1* for Participant-B. The total score for *H1.1* was then calculated as the mean of all participants. Continuing with the example, we would have calculated '3' (i.e. (4+2)/2) as the final score for *H1.1.*

In our statistical tests we checked (via one-sample Student t-test) whether the general means of each hypothesis could be assumed to be greater than '4'. 'Four' is the midpoint of the 7-point Likert scale, so answers between 4 and 7 can be considered as positive answers. The t-value and p-value for all hypotheses are depicted in Table 4.9, and the detailed calculation is available in the supplementary material.

Responses to open-ended questions When discussing the perceived advantages of the method, four participants emphasised the innovative approach of GO-Plan, three acknowledged its flexibility, and two highlighted its user-friendliness. On the other hand, when considering disadvantages, five participants pointed out the method's complexity, while two expressed concern that GO-Plan could potentially slow down the FAIRification planning process. Furthermore, two participants highlighted the challenge of identifying their own role within the FAIRification project. One participant noted: "I think it is important to understand your role as a planner; i.e the

Table 4.9: Results of the calculated mean for each hypothesis, minimum and maximum value, associated standard deviation, t-value and p-value. In the one-sample t-test, we checked whether the mean of the results for each hypothesis was greater than 4 (the midpoint of the Likert scale).

Hypothesis	**Mean**	**Min**	**Max**	**Standard Deviation**	**T-value**	**p-value**
H1.1	5,083	4.000	6.000	0.736	3,606	0.008
H1.2	6,083	5.000	7.000	0.736	6,934	<0.001
H1.3	5,917	5.000	6.500	0.585	8,032	<0.001
H1.4	5,417	4.500	7.000	0.970	3,576	0.008
H1.5	5.833	3.000	7.000	1.602	2,803	0.019
H1.6	5,389	3,667	7.000	1,084	3,140	0.013
H2	4,167	3.000	5.500	1.033	0,395	0,354
H3	5,917	5.500	6.500	0.376	12,474	<0.001
H4	5,167	4.000	7.000	1,169	2,445	0.029
H5	5,667	4.000	7.000	1,211	3,371	<0.001
H6	6,167	5.000	7.000	0.753	7,050	<0.001

profile you have to assume."

This tutorial-based validation aimed to address the threats to validity identified in the first application to real-world case validation. In the tutorial, we included users with varying levels of expertise, applied the method to a different domain, and thoroughly validated GO-Plan using a formal approach to capture users' perceptions of the method, rather than comparing it to *ad hoc* FAIRification.

As shown in Table 4.9, there are indications that hypotheses *H1.1* to *H1.6* and *H3* to *H6* are valid ($p < 0.05$). However, there is indication against H2 ($p = 0.354 > 0.05$). We suspect that this could be attributed to the lack of prior experience among most participants with FAIR and FAIRification. Since the entire subject matter was unfamiliar to them, it may have confounded their perception of GO-Plan as complex or difficult to use when compared to the inherent complexity of FAIRification. Conversely, more experienced participants who are already acquainted with FAIRification might see GO-Plan as a tool that helps reduce this complexity.

Threats to validity

Building on Wohlin *et al.* [129] common types of threats to validity, we identified the following threats to this validation task:

Internal Validity: We acknowledge that the low number of responses to our questionnaire might have biased our observations and results. This bias is further compounded by the fact that most participants had a similar experience level wtih FAIRification (i.e. no to limited experience) and profile (most were ontologists and conceptual modellers). Therefore, it is necessary to conduct this experiment with a larger and more diverse group of participants.

Another potential threat is that the use of cards mimicking potential obstacles and the documentation template might have biased participants' perception of the method. To mitigate this, we based the cards on obstacles we encountered in real-world experiences. Despite the risk of bias, we argue that the cards contributed to making the scenario more realistic, a benefit we believe outweighs the risk.

Finally, another threat to internal validity might arise if participants did not fully understand the introductory presentations (e.g. on FAIRification and GO-Plan) or the mock case scenario description. To address this, we included Q&A sessions during the presentations to allow participants to clarify any doubts.

External Validity: This threat can also arise from the low number of participants with homogeneous profiles, making it difficult to assume that our results comprehensively represent reality. Additionally, a missing aspect from the real-world setting was that participants could not consult with experts on FAIR to verify their decisions or discuss possible solutions. For instance, one tutorial participant mentioned "not having domain experts to question" as a downside of the hands-on session.

Construct Validity: In the responses to the questionnaires, a participant mentioned that it was difficult to identify their own roles during the method's execution. We recognise this as a construct validity threat, as it might have been challenging for them to respond to the questionnaire while taking multiple roles into consideration (e.g. domain expert, modeller, FAIR expert).

Besides, TAM was used in this validation task as a guiding framework. Ideally, the model should be statistically tested as a whole. However, such an evaluation requires a sample size of at least 200 responses, as recommended by previous studies [120, 133]. However, given the logistical challenges of organising conference tutorials, such as the limited number of attendees, we expected a smaller turnout for our tutorial sessions. Despite this expectation, we decided to proceed with our approach due to its inherent advantages. Firstly, conducting a tutorial in a multi-disciplinary conference allowed us to raise awareness of effective FAIRification planning. At the same time, it facilitated the engagement of participants possessing a baseline knowledge in relevant areas like data management, semantic technologies, and conceptual modelling. Secondly, the use of TAM provided a structured and robust framework for designing the tutorials, collecting information from participants and systematically analysing the data. It also allowed us to formulate hypotheses based on well-established theories from the literature. Nevertheless, the limited sample size still allowed us to conduct a simplified statistical analysis of the model so to make our results more meaningful.

Conclusion: Similarly to the previous threats, conclusion threats to validity could also arise from the low number of respondents to the questionnaire and from the homogeneous participants profile. This is because it is challenging to draw general conclusions from limited data, even when these conclusions are guided by formal methods such as TAM. Therefore, taking this limitation into account, we only present our results as indications of the validity of our hypothesis, and not as conclusions to such.

In summary, the aspects that could not be mitigated in this validation task are: (i) a low number of participants, (ii) participants with similar backgrounds, (iii) the absence of experts on FAIR for participants to consult with during the hands-on sessions, and (iv) a lack of well-defined roles for each participant during GO-Plan's application. We attempted to address these issues in a second validation task, which is described next.

4.4.3 Validation in the second tutorial

The third validation task involved repeating the "insights in FAIRification planning tutorial" as part of the 15th International Semantic Web Applications and Tools for Health Care and Life Sciences Conference (SWAT4HCLS 2024) [134].

Methods and Materials

For this validation task, we reused the same methods and materials from the previous tutorial, but included minor modifications to address previously identified threats to validity. Firstly, we invited other experts on FAIR to accompany the tutorial execution. Each expert observed a group and provided consultancy when requested. Experts were instructed to be as unobtrusive as possible, allowing groups to remain independent. Secondly, we included instructions for participants to organise themselves into distinct roles during the hands-on activity and maintain these roles throughout the tutorial.

Participants were also encouraged to choose roles aligned with their own expertise. We suggested roles as defined in Table 4.2.

Results and discussion

In the second tutorial, we collected responses from 12 participants. The participants' backgrounds included ontologists, conceptual modellers, linked data specialists, data stewards/managers, information technology (IT) specialist, data warehouse specialist and project managers. Five participants had previously been involved in two or more FAIRification projects, with one of them having participated in over 10 projects. Three participants had experience with one FAIRification project, while the remaining three were entirely new to the topic.

Responses to close-ended questions Table 4.10 describes the answers to close-ended questions. The results were calculated using the same method described in Section 4.4.2.

Table 4.10: Results of the calculated mean for each hypothesis, with associated t-value and p-value. In the one-sample t-test, we checked whether the mean of the results for each hypothesis was greater than 4 (the midpoint of the Likert scale).

Hypothesis	Mean	Min	Max	Std. Deviation	T-value	p-value
H1.1	6,227	5.000	7.000	0.684	10,796	<0.001
H1.2	6,045	4.000	7.000	1.106	6,135	<0.001
H1.3	5,773	5.000	7.000	0.847	6,938	<0.001
H1.4	6,182	4.500	7.000	0.874	8,281	<0.001
H1.5	5.909	4.000	7.000	1.136	5,573	<0.001
H1.6	6,030	5.000	7.000	0.809	8,316	<0.001
H2	5,500	5.000	6.500	0.548	9,803	<0.001
H3	6,445	5.500	7.000	0.611	13,334	<0.001
H4	5,727	4.000	7.000	1.191	4,811	<0.001
H5	6,182	4.000	7.000	0.982	7,372	<0.001
H6	6,273	5.000	7.000	0.905	8,333	<0.001

Responses to open-ended questions When discussing the perceived advantages of the method, participants mentioned GO-Plan's flexibility, user-friendliness and comprehensiveness. One participant stated that the method “forces everyone involved to really understand what is going on”, and another wrote that GO-Plan covers “all aspects that are important to take into account”. A third participant mentioned that the method supports organising concerns in different steps, and another participant wrote: “I think FAIR is very abstract and [GO-Plan] helps users identify these steps. It also creates a good way to argument with project managers and domain experts why we are doing certain tasks and reduce communication issues.”

In terms of perceived disadvantages, seven participants mentioned the method’s complexity as a downside. One participant mentioned that “some parts are not easy for inexperienced [people].” Additionally, two participants mentioned the method’s slowness as a disadvantage, and one criticised the lack of separation between data and metadata specific steps.

When analysing Table 4.10 we can observe an indication that the mean values for all hypotheses are greater than ‘4’. This is supported by the significantly small p-values ($p < 0.001$). As a result, there is evidence that *H1* to *H6* are valid in the context of our evaluation. This suggests that participants recognise the benefits of goal modelling within GO-Plan. Furthermore, it indicates that they find GO-Plan easy to use and useful, which leads to a positive attitude towards using it in their daily practice.

When comparing the participants background from the first and the second tutorial, it is possible to observe an increase in the number of representatives from different backgrounds related to FAIRification, as illustrated by Figure 4.12. We consider this a positive impact on the results of the second tutorial. Achieving more diversity was desired, as it makes our results more representative and generalisable.

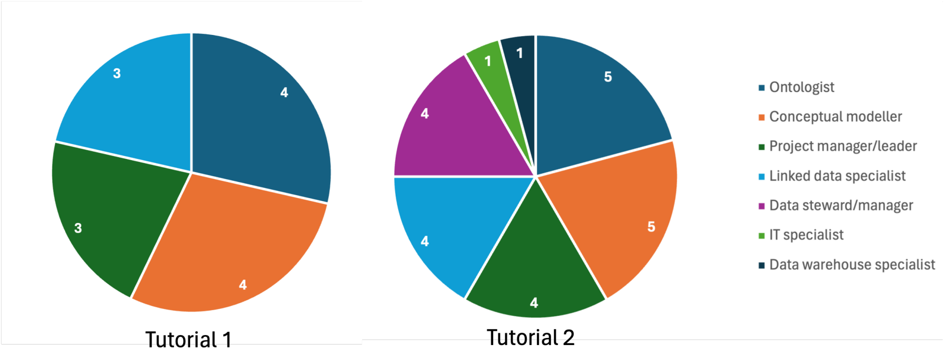


Figure 4.12: Comparison of participants background from Tutorial 1 (left hand side) and Tutorial 2 (right hand side).

Although the two tutorials were considerably similar, we decided to not merge their data because of the modifications introduced in the second tutorial. When comparing the results to close-ended questions from both tutorials, it is possible to observe a slight overall increase in the average of the hypotheses, as illustrated in Figure 4.13. Several factors might explain this improvement. First, a larger number of respondents in the second tutorial could reduce the influence of outliers. Second, the modifications implemented in the second tutorial may have contributed positively to the results. Third, the diversity of expertise among participants in the second tutorial, compared to the more homogeneous group in the first, might have played a role. Finally, the conferences themselves may have attracted different types of attendees, with the second conference being more relevant to the FAIR domain. These factors will be taken into account when planning future tutorials.

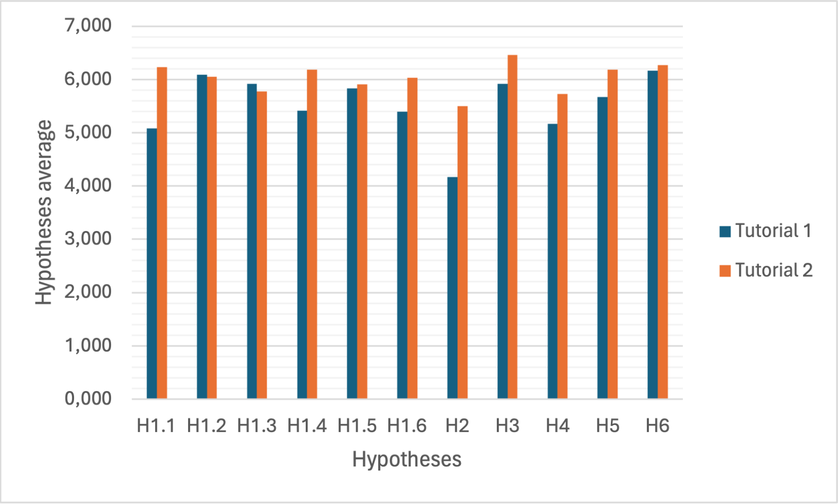


Figure 4.13: Comparison of the averages of each hypothesis calculated from the responses from the first and second tutorial.

From the participants' answers, we still observe significant concerns regarding the method's complexity and slowness. Notably, the participants who mentioned the method's complexity were those with experience in zero, one, or two projects. We consider this as an initial evidence to our argument that less experienced individuals may confuse the method's complexity with the inherent complexity of FAIRification. However, any conclusions on this matter require further investigation.

Threats to validity

In the second tutorial, we attempted to mitigate most of the threats identified during the first tutorial. However, as we could not significantly increase the number of participants, the threats related to the limited data still persist, albeit with a reduced impact.

It was also expected that some mitigation measures could introduce other threats to validity. For instance, the support of experts on FAIR could have biased some groups perception of GO-Plan (internal validity). To address this, we instructed experts to intervene only when requested and to remain as neutral as possible.

Another threat to validity could be related to the differences between the conferences at which the first and second tutorials were organised (construct validity). The first conference focused primarily on enterprise computing, while the second focused on semantic web technologies for life sciences. It is possible that one audience had more affinity with the topic of FAIR than the other. Nevertheless, it was important to test GO-Plan with different audiences. We intend to understand this impact in more depth by organising future editions of the workshop for other types of audiences.

4.5 Related works

Given the absence of existing literature presenting methods specifically tailored for FAIRification planning, we provide an overview of two distinct categories of related works: (i) resources offering guidance on FAIRification, such as FAIRification workflows; and (ii) works employing goal-based techniques to address data or data-management related tasks, such as designing a data governance plan. The intersection of these two categories represents a considerable part of the research areas that influenced the design of GO-Plan.

FAIRification guidance Several workflows and frameworks have been proposed to support FAIRification in different ways [114]. The generic [2], the *de novo* [35], and the FAIRplus [113] FAIRification workflows define the steps to be followed in the FAIRification of different types of FAIR resources, and they all describe the identification of FAIRification objectives as the first step of FAIRification. The FAIRplus FAIRification framework also includes a work plan layout to support organising the FAIR implementation work. The first phase of this framework consists of setting “realistic and practical goals” [113]. In this phase, useful recommendations and examples are provided with focus on defining an acceptable "FAIR enough" state for the resource to be made FAIR. A valuable recommendation given by FAIRplus is to avoid "the word 'FAIR' and its derivatives in goals entirely as it is too general to impart clear meaning" [113].

While these and other FAIRification workflows define a step for identifying FAIRification objectives [114], to the best of our knowledge, none of them have provided detailed guidance on defining FAIRification objectives or other FAIRification planning related aspects, such as distinguishing between the different types of stakeholders involved in FAIRification projects.

Goal-based data management or design Some works suggest the use of goal modelling to address various aspects of data, such as database design, data warehouse implementation, and data governance. For example, Jiang *et al.* [135] present a goal-oriented methodology for analysing and designing database requirements. The paper highlights that goal-oriented approaches capture not only the meaning of the data, but also who wants them and for what purposes, thereby providing additional information to support data integration and management. The proposed process starts with a step for capturing the needs of multiple stakeholders, which is followed by an analysis of alternative data requirements, thus resulting in a detailed conceptual schema for the data to be stored.

The work of Giorgini *et al.* [136] comments that a significant percentage of data warehouses fail to achieve their intended goals due to design issues. Their suggested method aims at analysing the high-level goals set by stakeholders and decision makers. Similar to GO-Plan, their method focuses on defining the goals of the collaborators and then identifying the essential data concepts required to support these goals.

Sothilingam *et al.*'s [137] work describes a goal-oriented approach to designing data governance. The paper highlights that goal modelling "facilitates the traceability of decision rationale and can help manage change over time. For example, one may want to re-visit past decisions and understand the logic which constituted those decisions." Their approach focuses on using goals as the primary guide for developing data governance schemes, and defines steps for modelling and evaluating goals. These steps include considerations such as alignment with business strategy (what the organisation wants to achieve), data strategy (how data will be managed), policy (enforcing compliance) and compliance plan (standards to be followed).

Although these methods share similarities with GO-Plan, such as the use of goal modelling and a focus on early requirements,[7] none of them concentrate on FAIRification. However, it is worth noting that some of these methods can complement the FAIRification process. For example, combining FAIRification with data governance is critical to ensure data quality. We argue that integrating data governance practices alongside FAIRification becomes essential to ensure the data adherence to the principles, but also to quality standards.

4.6 Final remarks

This paper described GO-Plan, a method created to support users in defining a FAIRification plan via the elicitation of FAIRification objectives. As the FAIR principles themselves, GO-Plan is neutral regarding the type of resource to be made FAIR, and therefore can be used to plan the FAIRification of various type of artefacts, including data and conceptual models (as demonstrated in the validations).

As mentioned in Section 4.3, real-world applications may benefit from using GO-Plan in an agile approach. In this case, the method can be fitted into one agile iteration and executed several times. It is up to the FAIRification team to decide how many iterations should be performed considering the project constraints (especially budget and time). Additionally, distinct FAIRification iterations can be tailored to address the specific needs and considerations of different stakeholders, thereby defining different levels of FAIR and related aspects for them. This is particularly valuable, for instance, when dealing with sensitive data (e.g. some types of users have access to different portions of data) or with FAIRification projects involving non-public data (e.g. from private companies), where certain reuse stakeholders might have limited access to (meta)data.

Furthermore, GO-Plan's modular structure allows for flexibility in its application, enabling users to selectively use its components to achieve specific aims. For example, an organisation undertaking a FAIRification project may find that phases 1 and 4 to 6 are sufficient, particularly if they have already conducted other FAIRification projects before. Users have the autonomy to adapt GO-Plan to their specific needs and circumstances.

Ultimately, GO-Plan builds on established methods from goal-oriented requirements engineering (e.g. [16]) and ontology engineering (e.g. [15]), allowing users to take

[7] In software engineering, *early requirements* refer to the initial stage of gathering and defining the needs, functionalities, and constraints of a software system or application. These requirements are typically identified and documented during the early phases of the software development life cycle.

advantage of their proven benefits. For example, the use of goal-modelling languages such as *iStar* can help stakeholders achieve a clearer understanding of objectives, make informed decisions, and improve communication [12, 132]. Moreover, other advancements in goal modelling, such as techniques for identifying risk (e.g. [138]), addressing obstacles (e.g. [139]), and resolving conflicts (e.g. [140]), can also be incorporated in the application of GO-Plan (e.g. in phases 5 and 6).

GO-Plan has been validated in a real-world application and in two tutorials conducted with participants of varied expertise. Also, its foundation on established methods contributes to the method's reliability, as these approaches have themselves been extensively validated. For example, Horkoff *et al.* [11] conducted an extended systematic study on goal-oriented requirements engineering that identified that 63.4% of papers have performed some kind of validation of their work (i.e. case studies, evaluation of scalability, controlled experiments, questionnaires, some type of benchmark), which shows that goal-based approaches have been extensively validated. Regarding the use of competency questions, Monfardini *et al.* [141] conducted a survey with 63 ontology engineers. Their results indicate that CQs have supported the definition of the scope and the evaluation of the ontology concepts. We argue that these findings also contribute to the robustness of GO-Plan.

The validation activities also indicate that GO-Plan meets the functional and non-functional requirements established during its design. The method has been tested in different domains (FR-1), such as the OntoUML catalogue and a dog shelter mock case, and applied to various resources types (FR-3), including a catalogue and mock data. Additionally, the method was tested on both types of FAIRification (FR-2): the catalogue case for retrospective FAIRification, and the mock case for both retrospective and *de novo* FAIRification. Based on the results from the two tutorials, we can assume that GO-Plan is relatively simple to use (NFR-1) and comprehensive enough (NFR-2). Although some users find it complex, there is evidence suggesting that, in general, the method is perceived as easy to use (i.e. hypothesis H2). Finally, users with varying levels of expertise stated that they were able to use the method effectively (i.e. H3).

We also contend that certain challenges from the research on FAIR fall outside the scope of GO-Plan. For instance, during the tutorials, one participant raised a discussion point about the blurred boundaries between what is considered metadata and what is considered data. They noted: "Very useful exercise, working on a toy example (but still quite detailed) such as this one helped me understand several aspects of FAIRification. I still believe one foundational question to be answered at the beginning should be 'what is my data and what is my metadata in this context.'

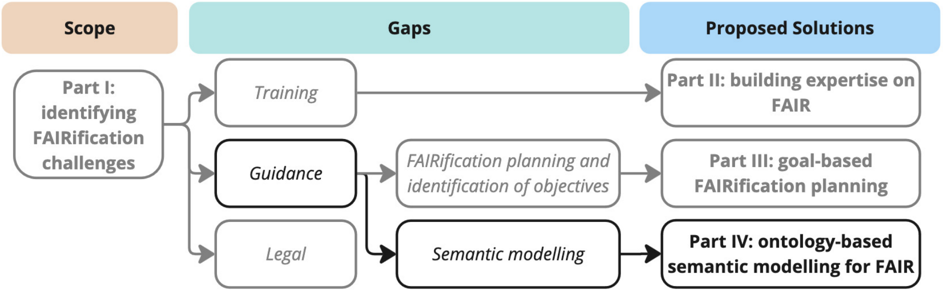
Indeed, while we were answering questions in phases 5 and 6 we often found confusion and disagreements in the group - derived by not having defined a stable separation on metadata and data.” This highlights that some difficulties are inherent to FAIR and FAIRification, or even to the broader semantic web community.

GO-Plan has not yet been validated in large projects (e.g. in multinational research environments) and this remains as a future work. Nevertheless, the method draws on prior experiences that involved such large-scale contexts (e.g. [8]). Additionally, although the primary focus of GO-Plan is to support small- to medium-scale institutions and research projects, we envision that larger projects could still be planned with the method if broken down into smaller parts.

The main aim of the work hereafter presented is to help all FAIR enthusiasts to better define clear FAIRification objectives that can lead to successful FAIRification. Nonetheless, we argue that communities should actively endeavour to share their FAIRification planning artefacts (e.g. goal diagrams, implementation decisions, FIPs) in order to promote standards convergence, disseminate solutions to implementation challenges, and share experiences so that others can prepare and execute FAIRification in a faster and more seamless way. To support this, we propose that FAIRification plans, including goals and mappings to related principles, should also be made FAIR. In addition to that, we emphasise the publication of FAIR implementation decisions (i.e. FIPs) as an effective means to gradually diminish the effort for subsequent projects and (re)users. In future work, we plan to continue to organise tutorial sessions similar to those described in the validation section. This ongoing effort is aimed at promoting awareness of proper FAIRification planning.

Part IV

Ontology-based semantic modelling for FAIR


Scope
Gaps
Proposed Solutions
Part I: identifying FAIRification challenges
Training
Guidance
Legal
FAIRification planning and identification of objectives
Semantic modelling
Part II: building expertise on FAIR
Part III: goal-based FAIRification planning
Part IV: ontology-based semantic modelling for FAIR

Topics covered in Part IV of this thesis.

The fourth and last part of this thesis explores how artefacts from the ontology and conceptual modelling research field can be adapted to the FAIRification context, which is connected to the need to provide better support for conceptual semantic modelling. Research in this area suggests that foundational ontologies can improve the quality of semantic models, under the assumption that they significantly support defining more interpretable and interoperable models. To investigate this, Chapter 5 presents a systematic literature mapping, examining 79 selected papers to understand the use and perception of foundational ontologies in biomedical research (the domain in which this thesis is contextualised). The aim of this chapter is to find evidence supporting or refuting the claimed benefits of using foundational ontologies. The analysis shows that foundational ontologies are used for various purposes, such as ontology construction, repair, mapping, and ontology-based data analysis. However, only one paper testing the claims about the benefits or drawbacks of using foundational ontologies in bioinformatics was found, and the experiments conducted in this paper were inconclusive.

Subsequently, Chapter 6 presents an experiment to test whether the advantages (e.g. clarity, consistency) of using ontology-based and well-founded semantic models hold true for machine learning-related initiatives. This experiment is motivated by the lack of empirical studies on the use of ontologies and semantic models in biomedical research, as identified in Chapter 5. Therefore, Chapter 6 examines the performance and consistency of predictions generated by a data-driven drug-repurposing pipeline. The experiment concludes that the use of conceptual models enhanced the reliability of ML predictions without compromising the ML model performance.

Although this work does not yet offer guidance on developing conceptual models for FAIRification, the findings from this part can inform future research on the topic.

For instance, the conceptual model presented in Chapter 6 was developed using a method in which goal models, such as those discussed in Part III, were employed to define the model's scope, indicating a correlation between the two.

Chapter 5

The use of foundational ontologies in biomedical research

César H. Bernabé, Núria Queralt-Rosinach, Vítor E. Silva Souza, Luiz Olavo Bonino da Silva Santos, Barend Mons, Annika Jacobsen and Marco Roos

Journal of Biomedical Semantics, 14, 21 (2023). https://doi.org/10.1186/s13326-023-00300-z

Abstract

Background The FAIR principles recommend the use of controlled vocabularies, such as ontologies, to define data and metadata concepts. Ontologies are currently modelled following different approaches, sometimes describing conflicting definitions of the same concepts, which can affect interoperability. To cope with that, prior literature suggests organising ontologies in levels, where domain specific (low-level) ontologies are grounded in domain independent high-level ontologies (i.e., foundational ontologies). In this level-based organisation, foundational ontologies work as translators of intended meaning, thus improving interoperability. Despite their considerable acceptance in biomedical research, there are very few studies testing foundational ontologies. This paper describes a systematic literature mapping that was conducted to understand how foundational ontologies are used in biomedical research and to find empirical evidence supporting their claimed (dis)advantages.

Results From a set of 79 selected papers, we identified that foundational ontologies are used for several purposes: ontology construction, repair, mapping, and ontology-based data analysis. Foundational ontologies are claimed to improve interoperability, enhance reasoning, speed up ontology development and facilitate maintainability. The complexity of using foundational ontologies is the most commonly cited downside. Despite being used for several purposes, there were hardly any experiments (1 paper) testing the claims for or against the use of foundational ontologies. In the subset of 49 papers that describe the development of an ontology, it was observed a low adherence to ontology construction (16 papers) and ontology evaluation formal methods (4 papers).

Conclusion Our findings have two main implications. First, the lack of empirical evidence about the use of foundational ontologies indicates a need for evaluating the use of such artefacts in biomedical research. Second, the low adherence to formal methods illustrates how the field could benefit from a more systematic approach when dealing with the development and evaluation of ontologies. The understanding of how foundational ontologies are used in the biomedical field can drive future research towards the improvement of ontologies and, consequently, data FAIRness. The adoption of formal methods can impact the quality and sustainability of ontologies, and reusing these methods from other fields is encouraged.

5.1 Background

Ontologies have long been used in biomedical research and applications [142]. For instance, these artefacts play an important role in improving the semantics and machine-actionability of Findable, Accessible, Interoperable and Reusable (FAIR) data and resources [1, 10]. Foundational ontologies are high-level, domain-independent ontologies constructed to provide basic categories and relations to concepts in domain-specific ontologies [14]. Theoretically, foundational ontologies are claimed to enhance the quality of domain specific ontologies and facilitate the interoperability among ontologies grounded on the same foundational one [14, 10, 143, 144]. However, it is difficult to find empirical evidence testing these claims.

The biomedical field has been developing and reusing tools to deal with the increasing growth in the volume of research data, which is impossible to analyse by human agents alone. As a result, several approaches have been proposed to make data and metadata (i.e., description of data) machine-readable and -actionable, such as to enable computers to understand and automatically process them (e.g., [1, 145]). To that end, the FAIR principles [1] focus on enabling efficient data analysis across multiple resources with minimal human intervention. The realisation of FAIR principles is intrinsically dependent on ontologies since they are used to describe, for instance, catalogues of resources (Findability), machine-readable access conditions (Accessibility), data and metadata (Interoperability and Reusability), and reuse conditions (Reusability) [32, 19].

In several fields, ontologies are used to model, represent, share and process knowledge about a domain. Ontologies became popular among bioinformaticians with the development of the Gene Ontology (GO) in 1998 [146, 147]. The success of GO has led many other groups to develop their own ontologies, which triggered initiatives such as the Open Biological and Biomedical Ontology (OBO) Foundry to coordinate ontology development efforts [147, 148]. To the current date, there are more than 250 active ontologies registered on the OBO Foundry Portal. Other ontology repositories also list an increasing amount of bio-ontologies. For instance, the NCBO Bioportal [149] catalogue currently contains more than a thousand ontologies.[1]

In general, the use of ontologies in the biomedical field faces several types of challenges. Some authors highlight the inherent diversity of the biomedical domain as one such challenge [150, 151]. This diversity can be perceived when capturing the domain's different levels of organisation, distinct types of entities, processes and rela-

[1] Information on the OBO Foundry Portal and the NCBO Bioportal checked on July 2023

tionships held by each entity. To illustrate, consider the multifaceted nature of proteins, which encompass sequences, functions, location, structure, interactions, related diseases, and so on [150]. This challenge is compounded by the constantly evolving nature of biomedical knowledge, which requires ontologies to continuously adapt as the field changes [150, 142, 151]. Additionally, ontologies play a significant social function by representing the collective knowledge and commitments of the communities that develop and use them [150, 142, 151]. As a result, there is a need for maintaining ontology quality, and fostering community awareness and acceptance of ontologies among all involved stakeholders (e.g., researchers, clinicians, developers, and end-users).

Furthermore, with the growing popularity of ontologies, it is possible to find different ones describing the same or a very similar domain scope. These overlapping ontologies can present conflicting definitions of the same concept, which impacts interoperability [152, 153]. To cope with that, prior literature suggests the organisation of ontologies in levels, where domain-specific (low-level) ontologies are grounded on domain-independent (high-level) ones, also known as foundational ontologies [153, 154, 155]. Most foundational ontologies reuse theories from cognitive science, philosophy, logic (i.e., description and first order logic [156]) and linguistics [14] to make clear philosophical distinctions about basic entities of the world [13, 157]. Arguably, these theoretic models can be used to articulate different conceptualisations across domains, and so, to enable interoperability [10]. Foundational ontologies are adopted by several research fields, including biomedical research [158]. Notably, the OBO Foundry defines a set of ontology development best practices that, for instance, proposes that each ontology should be built reusing a foundational ontology (more specifically, the Basic Formal Ontology (BFO) [159]).

The use of foundational ontologies can play an important role in the foreseen world of machine-actionable FAIR data, and also support bio-ontologists in dealing with the aforementioned challenges. However, we argue that this role must be well understood, and its expectations should be supported by empirical evidence. To address such need, this paper describes a systematic study of the literature to understand how foundational ontologies are used in biomedical research and its applications (including related and sub-fields such as bioinformatics). We seek to find empirical evidence supporting the claimed (dis)advantages of using foundational ontologies. Due to an apparent lack of adherence to formal methods in the development and evaluation of ontologies [160], we also explore how biomedical ontologies (developed using foundational ones) are developed and evaluated. Our approach is based on a Systematic Literature Mapping (SLM), which is a method to analyse the state-of-the-art on a particular topic [161].

5.1.1 Ontologies and ontological levels

Gruber [162] defines ontology as an "explicit specification of a conceptualization". This definition is extended by Studer, Benjamins & Fensel [163], who define ontologies as a "formal, explicit specification of a shared conceptualization". 'Formal' means that the conceptual model is logically defined so it supports algorithmic reasoning. 'Explicit' refers to concepts being defined with unambiguous descriptions. Finally, 'shared conceptualization' refers to the consensual definition of domain concepts within the community of expected users.

Usually, ontologies are organised in the application, domain, core and foundational levels [154]. Application ontologies are built to address a specific use case, usually constrained to a particular activity (e.g., orchestrate a machine learning workflow). Domain ontologies describe concepts related to a domain of discourse (e.g., rare diseases). Core ontologies provide an upper-level structural definition of a field that spans across different domains (e.g., biomedicine). Foundational ontologies define high-level (general) and domain-independent concepts (e.g., *process*, *quality*, *object*) that are articulated to define lower-level (more specific) ones. For example, the concept of *mitotic cell cycle* in GO can reuse the basic properties of *process* from BFO, thus inheriting its properties of having temporally proper parts and dependence on some material entities [159].

Some authors suggest that the ontological level should be seen as a continuous scale [155]. Figure 5.1 exemplifies the view proposed by de Almeida Falbo *et al* [155]. In this scale, the boundaries between foundational and core, and between core and domain/application ontologies are not always clearly defined. For instance, BFO is always considered a foundational ontology, positioned on the leftmost side of the continuum (exact classification). On the other hand, the Semanticscience Integrated Ontology (SIO) [164] and Biotop [165] are positioned between foundational and core ontology (our suggestion), since they define both domain-independent but also biomedical related concepts. The Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) [166], the General Formal Ontology (GFO) [167], the Suggested Upper Merged Ontology (SUMO) [168] and the Yet Another More Advanced Top-level Ontology (YAMATO) [169] are also examples of foundational ontologies (leftmost side of the scale, exact classification). The Orphanet Rare Disease Ontology (ORDO) [170] and the EJP RD CDE Semantic Model [20] can be seen as examples of domain and application ontologies (our suggestion), respectively. In the literature, foundational ontologies are also termed as "top" or "upper level" ontologies, while core ontologies can be named "domain upper ontologies" or "middle level ontologies".

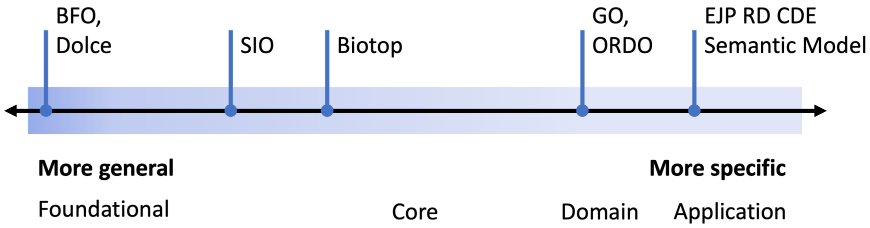


Figure 5.1: de Almeida Falbo *et al.*'s view of ontological levels as a continuum. Ontologies that define more general concepts of the world (e.g., foundational ontologies) are placed more to the left, while domain specific ones are placed more to the right. In the examples, BFO is defined in the leftmost side, while Biotop is depicted as more specific than BFO, but still more general than domain ones, such as ORDO. The positioning of SIO, Biotop, GO, ORDO and the EJP RD CDE Semantic Model are examples suggested by the authors of this paper.

5.2 The systematic literature mapping

The SLM research method used in this literature study is defined by Kitchenham [161] as a secondary study designed to answer broad questions about a research area. Planning, Conducting and Reporting are the main steps of the SLM process and are further divided into more specific tasks (Figure 5.2). In the planning step, the research questions, research sources, research query and selection criteria are defined. In the conducting step, papers are extracted from considered sources, deduplicated and selected according to inclusion and exclusion rules. In the reporting step, the results from the SLM are compiled and discussed in the form of answers to research questions.

Planning. We defined five research questions (Table 5.1) in the first task of the planning step. First, we wanted to know "How are foundational ontologies used in biomedical research?" (*RQ1*). Secondly, we wanted to investigate the reason why foundational ontologies are or are not used, and hence we asked two questions: "What are the claimed advantages of using foundational ontologies in biomedical research?" (*RQ2*) and "What are the claimed drawbacks of using foundational ontologies in biomedical research?" (*RQ3*). Third, since the answers to *RQ2* and *RQ3* are based on perceptions by the extracted papers' authors, we wanted to find scientific support for the answers to the questions by asking "What is the empirical evidence for the advantages and drawbacks of using foundational ontologies in biomedical research?" (*RQ4*). Finally, our second observation could be answered by asking "From the total

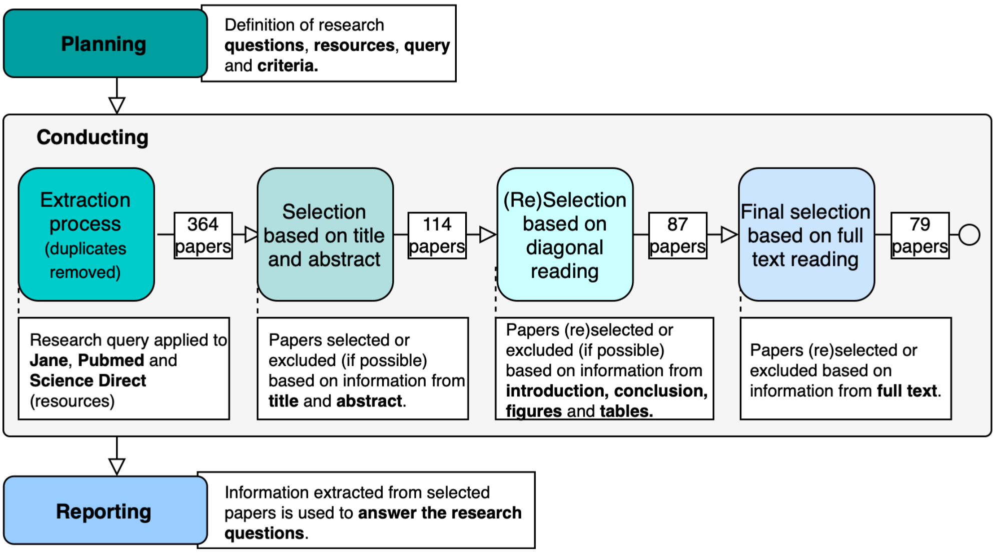

Planning
Definition of research questions, resources, query and criteria.
Conducting
Extraction process (duplicates removed)
364 papers
Selection based on title and abstract
114 papers
(Re)Selection based on diagonal reading
87 papers
Final selection based on full text reading
79 papers
Research query applied to Jane, Pubmed and Science Direct (resources)
Papers selected or excluded (if possible) based on information from title and abstract.
Papers (re)selected or excluded (if possible) based on information from introduction, conclusion, figures and tables.
Papers (re)selected or excluded based on information from full text.
Reporting
Information extracted from selected papers is used to answer the research questions.

Figure 5.2: Steps of the SLM method. The first step - planning - consists of defining the research questions, sources, queries and selection criteria. During conduction, the queries are applied to research sources, and the extracted papers are deduplicated and selected accordingly to different excerpts of information. In the last step, the selected papers are used to answer the research questions defined in the first step.

number of papers that describe the development of a biomedical ontology, how many use existing formal development and evaluation methods?" (*RQ5*).

ID	Research Question
RQ1	How are foundational ontologies used in biomedical research?
RQ2	What are the claimed advantages of using foundational ontologies in biomedical research?
RQ3	What are the claimed drawbacks of using foundational ontologies in biomedical research?
RQ4	What is the empirical evidence for the advantages and drawbacks of using foundational ontologies in biomedical research?
RQ5	From the total number of papers that describe the development of a biomedical ontology, how many use existing formal ontology development and evaluation methods?

Table 5.1: Research questions raised in the planning step of the SLM process. The first column identifies the research question, while the second column describes the question itself. The research questions are based on the literature.

Based on our research questions, we selected papers that discuss or use a foundational ontology in biomedical research related domains (Inclusion Criterion - **IC**). We excluded papers that did not mention a foundational ontology at all or were not related to biomedical research (Exclusion Criterion 1 - **EC1**). For the sake of reproducibility of this study, we also excluded papers not written in English (**EC2**). Table 5.2 summarises the inclusion and exclusion criteria.

	Inclusion criteria
IC	The paper discusses or uses a foundational ontology in biomedical research.
	Exclusion criteria
EC1	The paper is not related to biomedical research or no foundational ontology was mentioned.
EC2	The paper is not written in English.

Table 5.2: Inclusion and Exclusion criteria defined in the planning step of the SLM method.

Sources were selected considering the biological and computational aspects of biomedical research. We included one biosemantics focused source (**Jane** [171]), one biomedical (**Pubmed** [172]) and a third that also covers areas from computer science (**Science Direct** [173]). Finally, the search strategy was driven by the fact that different terms are used to describe foundational ontologies, as mentioned in the previous section. Thus, we included the "top level" and "upper level" synonyms in the search string, which is generically described in Table 5.3. Some small adjustments had to be done to fit the research string to the sources. For instance, it was not necessary to use the second part of the search string (the one after "AND") when extracting papers from Jane, since it is already constrained to the biomedical research field. The specific search strings, the search results (and the date each one was performed), are available in the supplementary material.

Search String
(”foundational ontology” OR ”top-level ontology” OR ”top level ontology” OR ”upper-level ontology” OR ”upper level ontology” OR ”upper ontology”) AND (”biology” OR ”biomedical” OR ”biomedicine” OR ”biological”)

Table 5.3: Search string defined in the planning step of the SLM. The search string is used to extract papers from the defined sources (Jane, Pubmed and Science Direct).

Conducting. In the extraction process, the search string was used in the mentioned sources and applied to the paper's full-text search. The search result was downloaded from each source, merged, deduplicated and selected according to the criteria defined. As described in Figure 5.2, the selection process is performed in three steps, where papers are first selected based on information from the title and abstract. Then, the results from the first step are reanalysed, now by performing diagonal reading (introduction and conclusion sections, figures and tables). In the third step, papers from step two are definitely selected/excluded based on full-text reading.

Reporting. In this final step, we compiled and reported the analysis conducted on the resulting selection of papers. During the final iteration of the selection process, the information from the papers was manually annotated and combined into mind maps that grew iteratively with each reading of a new paper. The mind maps and the information about the extraction process (which criterion was applied to each paper in each phase) are available as supplementary material.

5.3 Results

As illustrated in Figure 5.2, the first step of the extraction process resulted in 426 records, which were then deduplicated, resulting in 364 papers. The final step of the extraction process resulted in 79 papers, which comprehend works published in the years from 2004 to 2021. Overall, the number of publications per year is considerably stable, with peaks of nine papers dating from 2011, 2012 and 2015. DOLCE was the most used ontology in the years from 2004 to 2008, while BFO gained prominence from 2010 onwards. Additionally, BFO is the most used foundational ontology among the set of selected papers (42 papers), followed by Biotop (20), DOLCE (17), GFO (8), SIO (3), SUMO (3) and YAMATO (1). It is important to notice that some works used more than one foundational ontology in combination. We consider that a foundational ontology is "used" when it is applied to one of the activities described in the answer to *RQ1*.

From the set of selected papers, 57 of them developed or applied ontologies to specific fields of biomedical research, such as diseases (e.g., [174]), epidemiology (e.g. [175]), and genetics (e.g., [176]). The remainder of them (22) discussed or reviewed philosophical or logical aspects related to foundational ontologies, such as part-whole relations (e.g., [177]), granularity (e.g., [178, 179]) and dispositions (e.g., [180]). Following the SLM steps, we analysed the set of selected papers to answer the five research questions previously defined.

5.3.1 Synthesised Responses to Research Questions

The answers presented next synthesise our results. Detailed information can be found in the supplementary material, and a summary of the answers that follow is depicted in Table 5.6. Answers to *RQ1* ("How are foundational ontologies used in biomedical research?") can be classified into four categories:

- **Ontology development:** in most cases, foundational ontologies were used as a starting point for ontology design, providing a set of basic categories for deriving domain concepts (top-down approach). Foundational ontologies were also used in bottom-up (existing domain concepts are anchored in foundational ones), or middle-out (hybrid) approaches. For instance, Jensen *et al.* [181] used BFO to design the Neurological Disease Ontology (ND) in a hybrid approach. The authors first defined high-level classes and core entities that represent the domain using a top-down method (e.g., 'disorder' is_a 'material entity', 'diagnosis' is_a 'generically dependent continuant'). Then, primary literature and other clinical knowledge sources were used to identify more specific terms that were then connected to the high-level classes in a bottom-up strategy (e.g., 'protein aggregate' is_a 'disorder', 'diagnosis of Alzheimer disease' is_a 'diagnosis'). In summary, BFO supported both the characterisation of high-level classes and the categorisation of lower-level ones according to their nature (e.g., 'independent continuant' vs 'dependent continuant'). Foundational ontologies also supported the development of ontology design patterns (e.g., [182, 183]). For example, Schulz *et al.* [174] used BFO and Biotop to develop a design pattern to support distinguishing the structural, dispositional, and processual aspects of pathologies. The resulting design pattern reused the classes material entity, disposition, and occurrent to articulate the different interpretations of the pathology concept.

- **Ontology analysis and repair:** The ontological categories, relationships, constraints, and axioms defined by foundational ontologies were used to identify and repair inconsistencies in domain ontologies and other informational artefacts (i.e., information systems, databases, information flow processes, or documents), or to perform analysis to identify opportunities for improvement (e.g., clarity, accuracy). For instance, domain ontologies were grounded on foundational ontologies, which allowed for the identification and correction of inconsistencies and improvement of automated inference. To illustrate, Hoehndorf *et al.* [184] used fragments of BFO, DOLCE and GFO to identify contradictory class definitions across different biomedical ontologies. The authors found inconsistencies when interoperating the term secretion from GO and UBERON (Uber-anatomy Ontology) by applying the process and material object concepts from foundational ontologies. While GO treats secretion as a process, UBERON refers to it as a material object, which makes them incompatible (i.e., disjoint classes) even though they share the same label. Another example includes the work of Pisanelli *et al.* [185], in which theories from DOLCE were articulated to demonstrate the polysemy of the term "inflammation". In this case, it was concluded that the "inflammation" term could be refined into physiological function, a characteristic portion of a body part, a clinical condition, and a diagnosis applicable to that condition.

- **Ontology merging and mapping:** Foundational ontologies were used to support merging or development of mappings between different ontologies, usually with the aim of improving interoperability. In this case, foundational ontologies acted as semantic bridges between domain concepts, hence optimising the mapping process. For instance, in the work of Brochhausen *et al.* [182], the OMIABIS (Ontologised MIABIS) and BO (Biobank Ontology) ontologies, which are both grounded on BFO, were merged using upper-level BFO classes to create the Ontology for Biobanking (OBIB). The upper-level classes helped distinguishing and organising concepts related to, for instance, process (e.g., specimen collection) and material entities (e.g., specimen).
- **Ontology-based data integration and analysis:** domain ontologies developed with a foundational ontology were used to perform data integration and analysis. First, by grounding the data on an ontology, it was possible to identify errors, organise the data and connect it with external sources (e.g., other ontologised databases). Second, by being efficiently curated and by using reasoning, the ontology-grounded data could undergo a more significant data analysis (e.g., using machine learning algorithms), which can identify hidden knowledge [186], or present useful results that support clinical decisions [187]. In many cases, the domain ontologies served as common semantic data models for integrating heterogeneous data spread across different sources. In this context, foundational ontologies were expected to improve the axiomatisation and clarity of domain ontologies and data. For instance, Martinez-Costa & Abad-Navarro [188] mention that the use of Biotop enriched the axiomatisation of the common data model developed in the context of their work, which allowed for an unambiguous integration of domain specific knowledge. Additionally, the authors observe that "taxonomic reasoning allows queries to be performed at different granularity levels."

Answers to the question related to the claimed advantages of foundational ontologies (*RQ2*: "What are the claimed advantages of using foundational ontologies in biomedical research?") can be grouped into two main categories: **improvement of data** and **improvement of ontologies**. Advantages related to data are: enhancing data consistency and interoperability (by grounding data into unambiguous and interoperable ontological terms) and improving queriability (so ontologised data can be queried using human-readable terms). This is exemplified in the work undertaken by Masuya *et al.* [189], which uses YAMATO in a top-level ontology-based implementation of the RIKEN integrated database of mammals, which imports from several public knowledge sources (e.g., Ensembl, MGI). According to the authors, the approach allowed for "a consistent and scalable body of information that is interoperable with the global informational whole based on semantic web technology". Additionally, they state that "the standardized data formulation provided from top- and middle-level ontologies reduces the labor cost of data management through the reduction of unevenness in the operations of individual databases."

Regarding the improvement of ontologies, foundational ontologies are claimed to:

- improve the semantic understanding of terms and avoid ambiguity (as in the "inflammation" example above);
- enhance reasoning and prevent errors by, for instance, using the axioms added by foundational ontologies;
- speed up ontology development, through the reuse of top-level categories and other ontologies grounded on the same foundational one;
- improve interoperability, based on the idea that ontologies that use the same foundational ontology are expected to interoperate easier;
- facilitate ontology maintainability by reusing categories from foundational ontologies.

Examples of works that mention the advantages of using ontologies include Burek *et al.* [190], which notes that the "use of a top-level ontology potentially leads to fewer errors in the curation and creation of domain ontologies, a better understanding of the biological concepts and the means for data and ontology integration." Along the same lines, Keet [191] writes that "using a foundational ontology with its generic categories of entity types and core relationships across subject domains can facilitate bio-ontology interoperation, it speeds up ontology development."

A convergence to a small set of similar answers was observed when asking *RQ3* ("What are the claimed drawbacks of using foundational ontologies in biomedical research?"). Some works mentioned the complexity brought up by the use of foundational ontologies as the main drawback. This complexity is perceived in the time spent understanding class descriptions and in the high level of familiarity needed with background philosophical theories. Some papers described the difficulty to evaluate the claimed advantages as a demotivation towards using foundational ontologies. Additionally, the number of papers explicitly mentioning drawbacks (6 papers) is relatively smaller than the number of papers explicitly mentioning advantages of using foundational ontologies (43 papers).

By way of illustration, Some *et al.* [192] discuss certain advantages and disadvantages of using foundational ontologies. One downside entails a "tendency towards more complexity, especially regarding nested axioms", in which "simplification steps might be necessary when the ontologies are used in large KGs graphs, the performance of which might be affected by overly complex OWL models." In terms of evaluation, Boeker *et al.* [193] points to some difficulties that are also applicable to the use of foundational ontologies: "Due to the complex nature of the ontology artefacts, their evaluation is inherently difficult and manifold", as defining a good ontology depends "on the objectives of the ontology under scrutiny, its philosophical foundations and the intention of the investigator."

Unfortunately, just one paper specifically presented an empirical assessment to test the claimed advantages and drawbacks of foundational ontologies (*RQ4*). Boeker *et al.* [193] conducted a controlled trial to test the hypothesis that "students who received training on top-level ontologies and design patterns perform better than those who only received training in the basic principles of formal ontology engineering." In the assessment phase, students were asked to solve problems related to different topics, producing a set of ontological models that were compared to a gold standard. However, "the experiment showed no significant effect of the guideline-based training on the performance of ontology developers." The authors argue that, due to limited methodology, "the study cannot be interpreted as a general failure of a guideline-based approach to ontology development."

The last question aims to assess the methodological rigour in the process of building and evaluating ontologies (*RQ5:* "From the total number of papers that describe the development of a biomedical ontology, how many use existing formal development and evaluation methods?"). A subset of 49 of the 79 selected papers developed a new domain ontology using foundational ontologies. From these, 39 designed the ontologies using OWL [194], and 10 described the ontologies using other languages such as UML [195] (e.g., [196]) and FOL [197] (e.g., [198]). When analysing the subset of 49 papers, we investigated how they addressed ontology engineering [15] and ontology evaluation [199].

Ontology Engineering. Six papers stated that ontologies were built following the OBO principles [200]. As shown in Table 5.4, among the 49 papers that developed ontologies, four used Ontology Development 101 (OD101) [201], two used OntoSpec [202], one used the eXtensible ontology development principles (XOD) [203], one used Good Ontology Design (GoodOD) [204], and one used a combination of methods (OD101 and Methontology [205]). Two papers reused Ontology Design Patterns (ODPs) [206] to develop their own ontology. The other 33 papers did not use any method (*ad hoc*).

Ontology Engineering Method or Guideline	**Number of Papers Using the Method or Guideline**
OBO Principles	6
Ontology Development 101 (OD101)	4
Ontology Design Patterns	2
OntoSpec	2
XOD	1
GoodOD	1
Methontology + OD101	1
Total uses of approaches	**16**
Ad hoc	33

Table 5.4: Description of the ontology engineering methods used in 16 papers among the selected ones. The first column describes the name of the method, while the second mentions the number of papers that followed the method.

Ontology Evaluation. As described in Table 5.5, twenty-one papers evaluated their ontologies by using them in real-world or simulated application scenarios, more specifically: in data integration experiments, to support the development of an information system, in data classification algorithms, in ontology mapping experiments, and in querying and text mining tasks. Three papers used Competency Questions [124] as verification activities. Four papers performed instantiation [15] as a validation step and 3 papers validated the ontologies with domain experts. Two papers evaluated the ontology both through use case scenarios and domain experts. One paper used the *oQual* method [207]. Other studies (15) have not mentioned any kind of ontology evaluation.

Ontology Evaluation and/or Validation Method	**Number of Papers Using the Method**
Application to use case	21
Ontology instantiation	4
Competency questions	3
Validation with experts	3
Application + validation with experts	2
oQual	1
Total uses of approaches	**34**
No mention of evaluation	15

Table 5.5: Description of the ontology evaluation approaches used in the studies. The first column describes the approach followed. The second column shows how many papers used the approach.

5.4 Discussion

In this study, both the answers and the lack of answers to the research questions can be seen as results. Regarding *RQ1* ("How are foundational ontologies used in biomedical research?"), we found that foundational ontologies are mostly used in activities related to the development, mapping and repair of biomedical ontologies. Some papers also explored the ontology-based data analysis. We also observed that foundational ontologies were used in these tasks to support dealing with the challenges related to the diversity and complexity of the biomedical domain. As mentioned in the results section, several authors used foundational ontologies during ontology development to improve the semantic understanding of terms, to avoid ambiguity, to facilitate ontology maintainability and to improve ontology interoperability.

In general, we hypothesise two possible causes of the perceived complexity of using foundational ontologies (*RQ3*). The first hypothesis is that foundational ontologies are developed with excessive complexity and can be simplified. The second hypothesis, which would refute the first one, is that foundational ontologies are complex by nature, being complex solutions developed to solve complex problems (e.g., the inherent diversity of the biomedical domain).

Additionally, we argue that the answers to *RQ2* and *RQ3* ("What are the claimed advantages/drawbacks of using foundational ontologies in biomedical research?") need to be examined further before drawing any conclusions, as they may be perceived differently by different people. One concern is related to the expertise of researchers mentioning the (dis)advantages of using foundational ontologies: are they ontology experts or biomedical experts trying to develop an ontology? Do these two kinds of

ID	Question	Summarised Answer
RQ1	How are foundational ontologies used in biomedical research?	Foundational ontologies have been used in the development of domain ontologies and design patterns, ontology analysis and repair, ontology merging and mapping, and ontology-based data integration and analysis.
RQ2	What are the claimed advantages of using foundational ontologies in biomedical research?	The advantages of using foundational ontologies can be classified into two groups: improvement of data and improvement of core or domain ontologies. The former includes enhancing data consistency, interoperability and queriability. The latter is related to the improvement of semantic understanding of ontological terms, reasoning, inconsistencies prevention, ontology interoperability, maintainability, and a faster development process.
RQ3	What are the claimed drawbacks of using foundational ontologies in biomedical research?	The drawbacks of using foundational ontologies are related to their complexity and the difficulty in evaluating their claimed advantages.
RQ4	What is the empirical evidence for the advantages and drawbacks of using foundational ontologies in biomedical research?	We identified only one paper that performed an empirical assessment of the use of foundational ontologies in a biomedical research-related setting. The experiment did not reach any conclusion due to limited methodology.
RQ5	From the total number of papers that describe the development of a biomedical ontology, how many use existing formal ontology development and evaluation methods?	A subset of 49 papers developed a domain ontology. Among those, 16 used an ontology engineering method from the literature, and 34 performed a certain type of ontology evaluation.

Table 5.6: A summary of the synthesised responses to research questions.

researchers perceive complexity similarly or differently?

Another concern pertains to the influence of various tooling, modelling approaches (i.e., bottom-up, middle-out, top-down), and representation languages (e.g., OWL, FOL) on the ontologists' perception of foundational ontologies. For instance, we observed that OWL was used to represent a significant number (39 of 49) of ontologies newly developed. Indeed, this is also noted in the work of Flügel *et al* [156], who mention that OWL is more popular with developers because of its relatively user-friendly learning curve compared to, for example, FOL. However, due to its limited expressiveness, OWL cannot convey many ontological differences that are studied by foundational ontologists. Therefore, we anticipate that ontology developers engaged with FOL, who typically operate within a more advanced complexity stratum, will likely perceive foundational ontologies as less complex artefacts in contrast to those working within the lower complexity tier of OWL. Nevertheless, despite being significant questions for the research on the use of foundational ontologies, assessing these aspects within a SLM is a challenging task, as it would be difficult to measure the domain complexity and the experience level of all authors of the 79 papers, since this information is usually not available. Future experimentation aimed at assessing the claimed (dis)advantages of using foundational ontologies should consider these hypotheses and aspects.

We see the lack of answers to *RQ4* ("What is the empirical evidence for the advantages and drawbacks of using foundational ontologies in biomedical research?") as the main finding of this study. We identified only one paper ([193]) that ran an empirical experiment to test the use of foundational ontologies in the development of an ontology in the biomedical domain. It is important to note that the lack of experiments in biomedical literature does not imply that the claimed (dis)advantages of using foundational ontologies are under- or overrated. Actually, it indicates a clear need for evaluating these claims within the biomedical domain and testing the extended benefits for its applications.

Works in other research fields performed evaluations to provide empirical evidence for using foundational ontologies, and their results might be generalised to biomedical research (e.g., [208, 143, 209]). For instance, to test the usefulness of foundational ontologies in ontology engineering, Keet [143] conducted an experiment where participants had to choose between developing a "Computer Ontology" from scratch or reusing DOLCE or BFO. The study concluded that advantages brought up by using foundational ontologies make up for the time spent getting acquainted with them.

Verdonck *et al.* [209] also experimented the use of foundational ontologies in ontology development. Their work tested the differences between traditional conceptual modelling and ontology-driven conceptual modelling, in which foundational ontologies were used. The authors found out that few differences (e.g., number of ambiguities and inconsistencies) were noticed when participants had to model simple aspects of a domain. However, significant improvements were perceived when participants modelled more complex scenarios. We hypothesise that Verdonck *et al.*'s finding can explain the results of Boeker *et al.*'s experiment (from *RQ4*) because the models compared in the latter might not have been of significant complexity.

Finally, in *RQ5* ("From the total number of papers that describe the development of a biomedical ontology, how many use existing formal ontology development and evaluation methods?") we investigated how ontologies were built and evaluated. Our results show that only 16 of 49 papers used a systematic ontology engineering method or a set of guidelines. Similarly, only 4 of 49 papers used a formal evaluation method (Competency Questions or *oQual*) despite testing the ontology with a general approach (i.e., application or use cases, validation with experts, querying or instantiation).

Although foundational ontologies are claimed to impose a certain level of rigour during ontology development and evaluation, these processes need to be supported by additional techniques [15]. We theorise that using ontology engineering and evaluation methods should be an important concern in the research and development of ontologies, and that evidence is needed to demonstrate their benefit for biomedical applications. These methods guide ontology designers, data stewards and bioinformaticians in defining aspects related to the quality of content and sustainability of ontologies and ontology-based conceptual models (e.g., continuous integration, maintainability, documentation), which consequently impact the long-term realisation of the FAIR principles. In addition, ontology evaluation intends to identify inconsistencies in the developed ontologies, which should improve interoperability. The evaluation using use case scenarios is necessary, and it was done by several papers, but it also needs to be planned and performed with considerable rigour [210] and preferably combined with different approaches. Other research fields, such as the computer science domain, acknowledge that using ontology engineering best practices improves ontology consistency [10]. To exemplify, the research on ontology-based software engineering has been reusing several approaches from its own area of computer science (e.g., agile methods [211] and goal-modelling frameworks [212]) in ontology development. As such, incorporating this formal rigour for biomedical research can have the added value of increasing the FAIRness of ontologies and ontologised data.

Simon *et al.* [213] also mention that there are understandable reasons for the *ad hoc* features of many biomedical ontologies (e.g., lack of systematic ontology engineering methods, the non-use a foundational ontology), and we agree with the author's point of view. Given the urgency to move from paper-based to digital systems, ontologists were forced "to make a series of uninformed decisions about complex ontological issues", which can be understood in the context of our work as the lack of empirical testing and formal rigour in ontology development. The author also mentions that ontologists have been tempted to seek immediate solutions to particular problems but, to avoid further *ad hoc* problems, we strongly do not recommend this behaviour in a semantically interoperable digital world. To facilitate the adoption of formal rigour and engineering methods in bio-ontologies development, and to make ontologised data FAIRer, we suggest that both the ontological and biomedical communities work in closer and synergistic collaboration. We may assume that the more the ontology development methods and standards convergence within a community, the better and more interoperable the ontologies will be. Better ontologies will in turn result in better analysis and reuse of FAIR data. The extent to which the application of these methods and standards will translate into benefits for biomedical research will have to be demonstrated.

5.4.1 Limitations of this review

We recognise that some studies may not have been included in our analysis due to two reasons: (i) a paper may have used a foundational ontology without explicitly mentioning it, or (ii) a paper may have used an ontology that is in a grey area between foundational and core ontologies, and was consequently not properly captured. Additionally, terminological problems in the search string or in the coverage of the databases of the electronic libraries may have led to missing important studies. These can be seen as a trade-off in using a method such as an SLM, since definitions (e.g., whether a foundational ontology was used or not) must be clearly stated so the process can be systematically repeated.

The possible bias in the selection of papers could also have an impact on the results of this SLM. To mitigate this bias, we conducted periodic meetings between co-authors to discuss and validate the preliminary results of our analysis.

Finally, as previously mentioned, certain aspects that could influence the perception of benefits and drawbacks on the use of foundational ontologies in biomedical research (e.g., authors' experience, domain complexity, design method) were not measured as they could not be assessed from the information available in the papers.

5.4.2 Recommendations for biomedical applications and research

We expect that the results and discussions presented in this paper will inspire and guide future research and applications of foundational ontologies in the biomedical field. Examples of future endeavours may include tools to support classifying biomedical concepts into foundational ontologies' classes, domain-specific modelling languages that include theories from foundational ontologies, and efforts to educate people on proper ontology design. In fact, examples of initiatives from other fields are already in place. These include the "BFO classifier" [214], which suggests BFO classes for domain-specific concepts based on the users response to a decision tree questionnaire, and OntoUML [18], which is a modelling language that facilitates conceptual model design while using embedded theories from the Unified Foundational Ontology (UFO) [18]. Additionally, foundational ontologies are already being taught in computer science academic courses (e.g., [215]), and introducing them into biomedical courses would be beneficial.

Finally, other research paths could investigate the benefits of foundational ontologies in areas where they have been little explored, such as in machine learning (ML) and explainable artificial intelligence (XAI) applications. For instance, researchers can investigate whether XAI algorithms perform differently if trained on data that is organised accordingly to a model grounded on a foundational ontology, when compared to the ones trained on unstructured data. Similar investigations are discussed by Amaral, Baião & Guizzardi [144] in their paper about the use of foundational ontologies for data mining. The authors argue that the quality of data mining results is related to the extent that they accurately reflect the real world, and add that the "fundamental ontological distinctions embodied in a foundational ontology can be used to improve the quality of the data mining process, mainly when it includes information from multiple sources that may commit to different theories about a particular concept."

5.5 Related works

The related works listed in this section have reviewed the literature to investigate the use of foundational ontologies in other fields. Nardi, Almeida & Falbo [216] performed a Systematic Literature Review (SLR) to study the use of foundational ontologies for semantic integration in Enterprise Application Integration (EAI) research. This research area focuses on the development and use of plans, methods, and tools to integrate distinct information systems. They identified that foundational ontologies have been used to solve semantic conflicts between the applications' concepts, to develop core or domain ontologies, and to integrate different ontologies and databases. The authors also described that most systems and ontologies were developed without any systematic approach (*ad hoc*).

Elmhadhbi, Karray & Archimède [217] investigated the role of foundational ontologies as a means for the formalisation and integration of heterogeneous resources for information systems. The authors concluded, based on the literature and their own experience, that the "use of upper ontologies improves data quality, reduces development time and especially facilitate large-scale information integration by avoiding ambiguities or inconsistencies to guarantee semantic interoperability of systems."

Baumgartner & Retschitzegger [218] presented a survey on the use of foundational ontologies for situation awareness, which is a research area that focuses on the decision-making process under complex and dynamic situations. The authors point out three types of uses of foundational ontologies in computational approaches to situation awareness: integration of heterogeneous information, identification of relevant situations in a domain-independent way, and knowledge sharing across domains.

Trojahn *et al* [219] performed a survey on the use of foundational ontologies for making domain ontologies interoperable. The work provides an overview of various ontology-matching activities that can benefit from foundational ontologies. They state that the potential of foundational ontologies for clarifying semantics enhances ontology quality, avoids poor ontology design and facilitates interoperability between ontologies. In regard to the challenges, they state that the "problem of matching ontologies gets more complex when involving foundational ontologies", as it "requires the deep identification of the semantic context, the identification of subsumption relations, and consistency with the formalization." They conclude that the main challenge in using foundational ontologies relates to the need of specialised knowledge, as using foundational ontologies demands a thorough understanding of their underlying philosophical theories.

Several other works in the literature have also conducted similar reviews and analyses (e.g., [218, 220]). Most of them, including ours, concur on the benefits (e.g., enhance interoperability and semantic clarity) and drawbacks (e.g., complexity) of using foundational ontologies. Likewise, they recognise similar challenges and requirements (e.g., the need for evaluation, systematic approaches and specialised expertise) for advancing the use of ontologies in their field of research. However, none of them have attempted to identify empirical experiments that test the claims of using foundational ontologies. To the best of our knowledge, no studies have reviewed the use of foundational ontologies in the biomedical research field so far.

5.6 Conclusion

This paper described a Systematic Literature Mapping conducted to understand how foundational ontologies are used in biomedical research and to identify the empirical evidence in favour or against claimed advantages. Additionally, we investigated the level of methodological rigour in papers that used foundational ontologies to construct domain ones. Understanding how foundational ontologies are used in biomedical research and applications can better drive future research towards the improvement of ontologies, and consequently the FAIRness of ontologised data. Our findings imply two main conclusions. First, there is a lack of empirical evidence in biomedical literature for or against the use of foundational ontologies. Second, this particular area of biomedical research does not apply ontology development and evaluation more formally and systematically. Consequently, we recommend that research in bio-ontologies addresses the creation or reuse of methods for ontology engineering (considering phases from ontology requirements elicitation to testing and sustainability) and ontology evaluation (encompassing both evaluation techniques and procedures for application-based evaluation) supported by foundational ontologies. Future research could investigate how foundational ontologies are benefiting biomedical applications, how they are used in other fields and what can be reused to improve research in ontologies for biomedicine.

Chapter 6

Restructuring knowledge graphs with conceptual models

Implications for machine learning predictions in drug repurposing

César Bernabé, Rosa Zwart, Pablo Perdomo Quinteiro, Annika Jacobsen, Núria Queralt-Rosinach, Katherine Wolstencroft, Luiz Olavo Bonino da Silva Santos, Barend Mons, Marco Roos

Advanced Information Systems Engineering. CAiSE 2025. Lecture Notes in Computer Science, vol 15702. Springer, Cham.
https://doi.org/10.1007/978-3-031-94571-7_17

Abstract

This paper investigates the impact of restructuring knowledge graphs (KGs) with well-founded conceptual models to improve machine learning (ML) predictions, particularly in drug repurposing applications. These conceptual models were developed using OntoUML, which is grounded in the Unified Foundational Ontology, and were constructed following an established workflow for data FAIRification–a process aimed at making data more Findable, Accessible, Interoperable, and Reusable. We compared the performance of a Graph Neural Network model trained on original public KGs with models trained on the same restructured KGs. Our results indicate that while the ML model classification performance (measured in terms of accuracy and error metrics) remains similar for both, the models trained on restructured KGs produce more consistent predictions, reducing variability across multiple runs. These findings suggest that restructuring KGs using well-founded conceptual models can enhance the reliability of ML predictions without compromising model performance. We conclude by proposing future research directions to further explore the potential of conceptual models and FAIR principles in improving ML.

6.1 Introduction and Background

Machine learning (ML) models are often trained using large knowledge bases [221]. However, constructing such voluminous datasets is both resource-intensive and time-consuming, as existing data is typically not prepared for reuse [222]. To facilitate data reusability, the FAIR principles were introduced to guide the process of making data and other resources Findable, Accessible, Interoperable, and Reusable [1]. Since their publication in 2016, the principles have gained significant traction across various fields [223]. Similarly, research and applications involving ML models have expanded rapidly in recent years [221]. However, although these areas complement each other, there has been little research on the specific impact of FAIR data on ML methods.

Jacobsen *et al.* [2] proposed a stepwise process of making existing data FAIR (referred to as FAIRification). The generic FAIRification workflow is organised in steps, starting with the identification of FAIRification objectives, followed by the analysis of (meta)data, the design of semantic models for (meta)data, and (meta)data restructuring, linkage, hosting, and assessment. In the semantic modelling phase, a conceptual model [224] of the data elements (e.g. patient, disease) and relationships (e.g. drug treats disease) is constructed. Since FAIR promotes reuse by both humans and machines, the conceptual model is designed to be as accurate a reflection of the data domain as possible [223]. In the data restructuring step, the data to be made FAIR is reorganised to align with the structure of the conceptual model, making it not only more understandable for humans but also more easily integrated with other FAIR data.

In this work, we aim to assess the impact of this conceptual model-based data restructuring on ML models. To conduct this experiment, we build on parts of the pipeline developed by Perdomo-Quinteiro *et al.* [225], which reuses data from public sources to create a knowledge graph (KG) for training a Graph Neural Network (GNN) to predict drugs that can be repurposed to treat symptoms of rare diseases. We replicate the data fetching process of Perdomo-Quinteiro et al.'s pipeline (named *rd-explainer*) to generate an initial KG. Then, we restructure the KG previously generated based on a conceptual model and compare the performance and outputs of the GNN model when trained on both cases.

Our results highlight promising directions for future research, despite being limited to a single domain and ML algorithm. Experimentation shows that models trained on the conceptual-model-based KGs (CM-based KGs) produced more consistent predictions (i.e. less random), with variability in predictions across different runs of the

GNN being 29.91% lower compared to those on the original KGs. Furthermore, the predictive performance of the GNN model trained on the CM-based KGs did not show a significant difference from the original KGs in terms of accuracy and error metrics. These findings suggest that further exploration of CM-based KGs could yield valuable insights. Indeed, Perdomo-Quinteiro *et al.* note that reproducibility is an issue that needs further investigation: "The known reproducibility issue of our explainer that may imply that the explanations are different each time it is used, may reduce the confidence and reliance on the explanations." [225]

For the sake of clarity, it is important to note that the term "model" carries multiple definitions depending on the research field. In the context of machine learning, a "model" is a mathematical representation or algorithm used to make predictions or decisions based on data [226]. Conversely, in conceptual modelling research, a "model" serves as a structured framework that represents the concepts and relationships within a specific domain, thus providing a formalised approach to organising and interpreting information [224]. To maintain clarity, we differentiate between these definitions using the wordings "ML model" and "conceptual model" to define the different interpretations in each case, respectively.

The remainder of this paper is organised as follows: Section 6.2 provides a brief overview of *rd-explainer*. Section 6.3 describes the method used in our explorations. Section 6.4 presents our results, followed by a discussion in Section 6.5. Sections 6.6 and 6.7 address the limitations of our study and related works, respectively. Finally, Section 6.8 concludes the paper.

6.2 The *rd-explainer* use case

The experimentation described in this work builds upon the method presented by Perdomo-Quinteiro *et al.* [225], which developed *rd-explainer*, an interpretable ML method for drug repurposing. Drug repurposing identifies new uses for existing drugs, a cost-effective strategy particularly valuable for rare diseases with limited pharmaceutical interest [227]. This technique has already been successfully employed in many health care domains, such as in cancer therapy (e.g. [228]) and rare diseases themselves (e.g. [229]). *rd-explainer* relies on aggregated data sourced from three key public knowledge bases: the Monarch Initiative [230], Drug Central [231], and the Therapeutic Target Database [232].

During the pipeline execution, data is initially retrieved from Monarch using a disease code (from a ontology) as the starting point for constructing the initial KG.

The fetching script uses the disease code to identify the corresponding disease node in Monarch, and subsequently fetches all nodes directly related to it. The KG is then enriched with data from Drug Central and the Therapeutic Target Database, incorporating information about drugs and their associated treated symptoms. The data from these two additional sources is adjusted to conform to the original graph structure defined by Monarch.

Subsequent to generating the disease-specific KG, a GNN model (GraphSAGE [233]) is trained on it. The output of this process is a ranked list of predictions, with each entry representing the probability of a link existing between a drug and the target symptom. The higher the score, the greater the likelihood of an actual edge existing between the two nodes, indicating a stronger potential relationship between the drug and the symptom or disease. For more information on *rd-explainer*, the reader can refer to Perdomo-Quinteiro *et al.* [225].

6.3 Method

To restructure the KG produced by the *rd-explainer*, we followed relevant FAIRification steps: identification of FAIRification objectives, data analysis, conceptual modelling, and data restructuring. Subsequently, we compared performance and results of the GNN model when trained on the original KG and on the CM-based one. An illustration of the method described in this section, along with the detailed FAIRification objectives, the original KG metamodel, the conceptual model, the data-fetching and training scripts, the performance measurements and resulting predictions are available in the supplementary material [234].

Identification of FAIRification objectives. The FAIRification objectives identified in this step focus on making data reusable for drug repurposing applications. These include 'identifying existing drugs that can be repurposed to treat the symptoms of rare diseases', as well as the sub-objectives 'identifying drugs that target genes associated (in)directly with a rare disease' and 'identifying drugs known to treat phenotypes associated (in)directly with a rare disease.'

Data analysis. In this step, the original KG constructed during the execution of *rd-explainer* was analysed, and its schema was extracted to be used as a starting point for conceptual modelling. An excerpt of this schema is illustrated in Figure 6.1, and its complete version is available in the supplementary material [234].

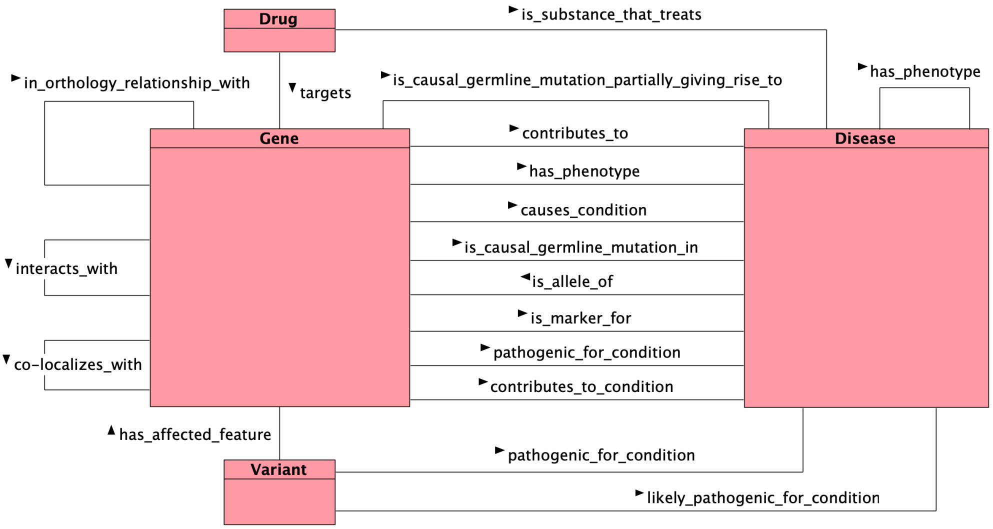


Figure 6.1: An excerpt of the original KG schema, showing the concepts Drug, Disease, Gene, and Variant. Rectangles denote node types (rdfs:Class), and lines denote relationship types (rdf:Property).

Conceptual modelling. The conceptual model for the drug repurposing domain was developed iteratively. This involved designing the domain model using Onto-UML [235], a modelling language based on the Unified Foundational Ontology (UFO) [18], which supports the creation of ontologically *well-founded conceptual models*, ensuring semantic clarity in representing real-world phenomena.

The resulting model, designed with FAIRification goals in mind, underwent validation through three rounds of expert review involving bioinformaticians and ontologists. The expert group comprised postdoctoral researchers and PhD candidates, with two university professors also participating in the initial session. All had prior experience in rare disease projects, and most had been involved in drug repurposing efforts for over three years. Each review session included, on average, eight participants. During sessions, the model was presented alongside the FAIRification objectives, and experts were invited to ask clarifying questions and assess the accuracy and relevance of the concepts and relationships. Suggestions were discussed collectively, and those reaching consensus were incorporated into the model. By the third round, the experts agreed the model was sufficiently robust and ready for use.

An illustration of the resulting conceptual model is shown in Figure 6.2. It represents key biological entities and their relationships within the domain of drug repurposing and (rare) diseases. At the core, the model involves entities such as Gene,

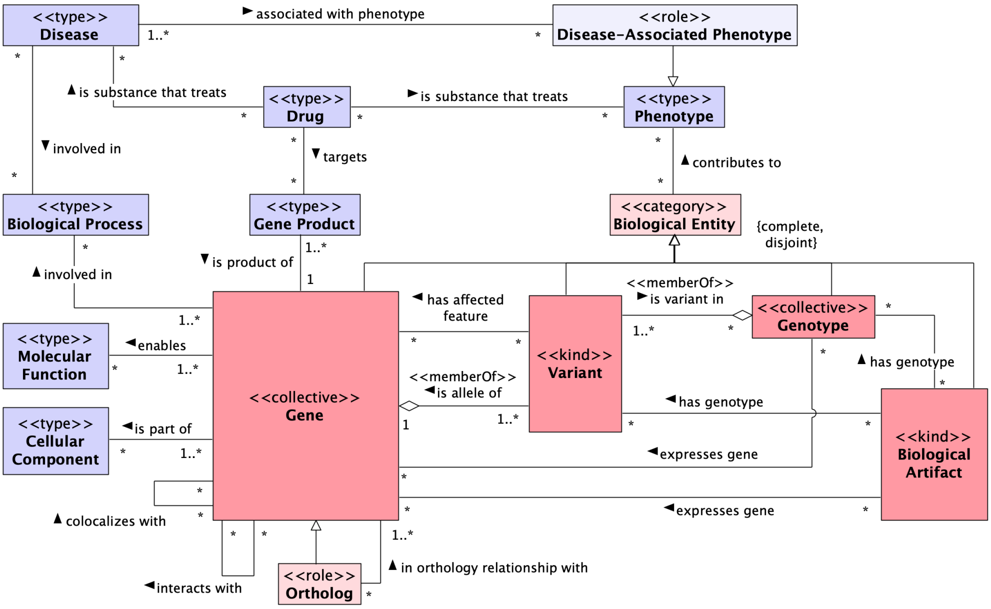


Figure 6.2: Conceptual Model for the drug repurposing domain. The resulting CM-based KG mirrors the structure of elements (nodes) and relationships shown in the model.

Variant, Genotype, and Phenotype, which are fundamental to understanding genetic and phenotypic expressions of diseases. A Gene is a collective biological entity that may have interactions (e.g. it interacts with or co-localises with other genes). An Ortholog, shown as a subtype of Gene, refers to genes in different species that evolved from a common ancestral gene by speciation and typically retain the same function, making them crucial for studying disease mechanisms across species. Variants are specific alterations in a gene, and they can be part of a Genotype, which represents the complete set of an organism's genetic information. The model shows how Variants and Genotypes *express* Genes, impacting Biological Processes like Molecular Functions and Cellular Components. Additionally, Drugs are connected to Diseases and Phenotypes through their treatment relationships (*is a substance that treats*), thus targeting Gene Products, which are produced by genes and influence disease-related functions.

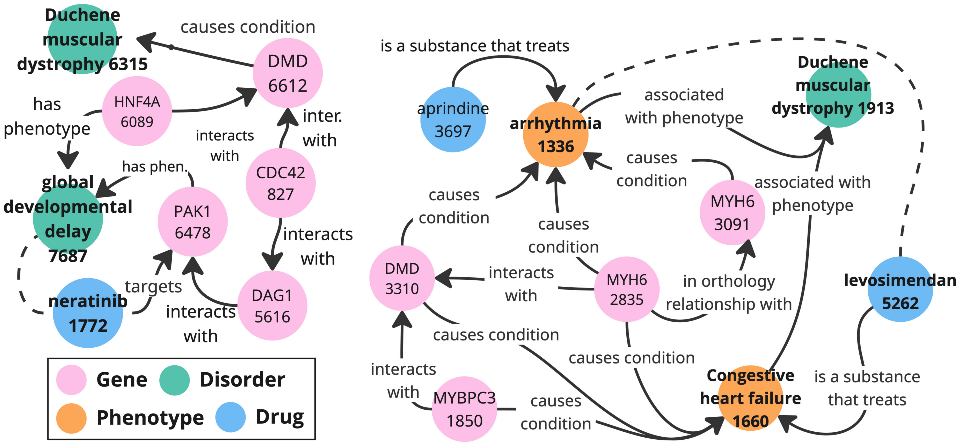
To exemplify the model in Figure 6.2, consider a rare neuromuscular Disease, XYZ syndrome, caused by a specific Variant in the Gene XYZ1. This Variant is part of the patient's overall Genotype. The variant results into an altered Gene Product that is no longer able to reach or function in its intended Cellular Component—the neuromuscular junction—where it normally plays a role in transmitting nerve signals. As a result, the corresponding Molecular Function is disrupted, leading to the Phenotype of *severe muscle weakness*. Research reveals that XYZ1 also has an Ortholog in mouse (xyz1), enabling scientists to model the effects of the variant in vivo and study the Biological Process underlying neuromuscular synapse formation across species. This helps, for instance, identifying a potential Drug through repurposing efforts: it binds to the dysfunctional gene product and partially restores its function, thus mitigating the Disease-Associated Phenotype and offering hope for clinical treatment of XYZ syndrome.

Data restructuring. Following the conceptual modelling step, the initial KG was reorganised according to the elements and relationships defined in the conceptual model from Figure 6.2. To achieve this, the data-fetching script from *rd-explainer* was modified to generate the CM-based KG. The mapping from the original to the CM-based KG was manually reviewed by one of the authors and an external bioinformatician. All data used in this study were retrieved from the September 2021 version of Monarch on 1 May 2024. Data from Drug Central and the Therapeutic Target Database were also collected on 1 May 2024.

GNN prediction and performance assessment. After constructing both the original and CM-based KGs, we proceeded to train separate GNN models on each KG type. This process was repeated ten times per KG type to collect performance metrics that were averaged to ensure a balanced comparison (i.e. AUC-ROC [236], which evaluates a model's capacity to distinguish between classes, where a higher score implies better classification; F1 score [237], which is the harmonic mean of precision and recall; and Cross Entropy Loss [238], which measures the difference between predicted probabilities and true labels, with lower values indicating a more accurate model performance). At the end of this process, 20 prediction lists are generated: 10 lists from the original KG and 10 from the CM-based KG.

Next, to assess the reliability of the ML models, we evaluated the consistency of their predictions. This evaluation was motivated by Perdomo *et al.*'s observation regarding the challenges of ensuring reproducibility in ML methods. For instance, given the stochastic nature of some components of *rd-explainer* (e.g. edge2vec [239]), running the same process multiple times can lead to different sets of predictions. Therefore, a reliable model should produce consistent results that are not heavily influenced by randomness. To measure this, we performed pairwise comparisons of the prediction lists separately for each of the two groups of 10 iterations from each KG type. For each pair of prediction lists, we calculated the percentage of overlapping predictions, with a lower overlap indicating less consistency across iterations and greater variability in the ML model's outcomes.

Targeting a rare disease. As previously mentioned, the process described above requires specifying a target disease (i.e. a disease code) when constructing the KGs, meaning the GNN models are trained on disease-specific KGs. To gather more comprehensive insights from our experiments, we applied our method to three different rare diseases: Duchenne Muscular Dystrophy (DMD) [240], Huntington's Disease (HD) [241], and Osteogenesis Imperfecta (OI) [242]. These diseases have been chosen to maintain consistency in disease characteristics, as each has a single causative gene, in contrast to diseases like Alzheimer's Disease and Amyotrophic Lateral Sclerosis, which were used in Perdomo-Quinteiro *et al.*'s experiments. This process resulted in the construction of six distinct KGs—two for each disease: one original KG and one CM-based KG—and enabled disease-specific comparisons.



(a) original KG

(b) CM-based KG

Figure 6.3: Illustrations of the knowledge graphs and associated predictions for DMD as target disease. The dashed line represents the prediction itself, while the remaining lines and nodes depict relationships that influence the prediction. Node types are distinguished by colour, as shown in Fig. 6.3a.

For illustration, Figure 6.3 shows an extract of the original and CM-based knowledge graphs generated for DMD as the target disease. It is important to note that assessing the correctness of these predictions is beyond the scope of this paper, as doing so would require proper drug evaluation procedures.

6.4 Results

A readily applicable outcome of this work is the conceptual model developed during our method (Figure 6.2). Moreover, our assessments indicate that the GNN models trained on both types of KGs achieved comparable performance in metrics such as AUC and F1 scores. More significantly, the comparison in terms of output reliability reveals that models trained on CM-based KGs produced more consistent predictions (i.e. similar prediction results). High-resolution versions of the figures presented next and the list of predictions are available in the supplementary material [234].

6.4.1 The OntoUML-based conceptual model is reusable

The conceptual model designed in this work can be reused in other ML systems and FAIRification processes, as well as be extended for other applications in related domains. When comparing the original and restructured KGs (Figures 6.1 and 6.2), it can be observed that the number of nodes increased from the original KG to the CM-based KG. In contrast, the number of relationships decreased significantly, as some concepts previously defined as relationships in the original KG were transformed into concepts in the restructured version. For instance, the *has phenotype* relationship in the original KG, which linked Gene and Disease, was transformed into a Phenotype concept in the restructured KG.

6.4.2 Predictive performance is similar

The (averaged) training curves of the GNN models are illustrated in Figures 6.4, 6.5, and 6.6, for DMD, HD, and OI, respectively. Figures 6.4a, 6.5a, and 6.6a display the training metrics of the GNN models trained on the original KGs, while Figures 6.4b, 6.5b, and 6.6b show the metrics from the training on the CM-based KGs. Each figure presents the AUC-ROC scores and the Cross Entropy Loss (CEL) of the training processes. When comparing the (a) and (b) versions of each figure, it is important to note that the training curves differ in the total number of epochs as distinct hyperparameter optimisation processes (random search) [243] were performed for each of the six KGs–this is motivated by our aim to evaluate the impact of CM-based restructuring on the entire process. The hyperparameters used in each case are described in the supplementary material [234].

Overall, for all models trained on both the original and CM-based KGs, the training process starts with a high AUC-ROC score for both the training and test sets. However, the improvement from the start to the end of training is minimal. The CEL curve exhibits varying behaviour depending on the target disease. For DMD, a more rapid decline in the CEL curve is observed when training on the original KG compared to the CM-based KG. In contrast, the opposite trend is seen for HD and OI, where training on the CM-based KG results in a steeper decline.

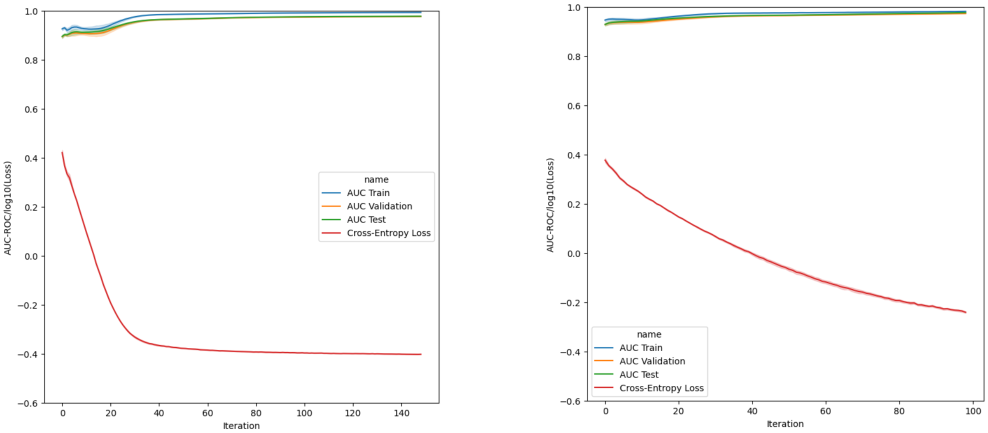
Table 6.1 shows a summary of the average AUC-ROC and F1 scores for each case and target rare disease. For DMD, both the AUC-ROC and F1 scores are slightly higher when training the GNN model on the original KG, although the difference is minimal. For HD, training on the original KG resulted in a higher average AUC-ROC, while training on the CM-based KG resulted in a higher F1 score. For OI, the average AUC-ROC and F1 scores were higher when the ML model was trained on the CM-based KG.

Table 6.1: AUC-ROC and F1 scores for training on original and CM-based KG, for each target rare disease.

Disease	DMD		HD		OI	
KG Type	Original	CM-based	Original	CM-based	Original	CM-based
AUC-ROC	**0.977**	0.976	**0.978**	0.967	0.9602	**0.974**
F1	**0.933**	0.906	0.896	**0.934**	0.812	**0.915**

6.4.3 Predictive consistency is higher for CM-based KGs

Figures 6.7, 6.8, and 6.9, illustrate the degree of overlap (expressed as a percentage) between predicted drug-phenotype pairs across all ten runs for the original and CM-based KG for DMD, HD and OI, respectively. Figures 6.7a, 6.8a and 6.9a are derived from the predictions of the GNN model trained on the original KG, whereas those in Figure 6.7b, 6.8b and 6.9b are derived from the predictions of the GNN model trained on the CM-based KG. The means and median of the overlaps described in Figures 6.7, 6.8, and 6.9 are summarised in Table 6.2. Given the mean and median values of the overlap of predicted drug-phenotype pairs, it can be observed a higher mean and median for when *rd-explainer* is applied to the CM-based KG in the experiments conducted for all target rare diseases.



(a) Original KG

(b) CM-based KG

Figure 6.4: Training curves of the GNN models using the original and CM-based KG as input, for **DMD**. The blue, green and orange lines depicted in the figures on the right are the variations of AUC values among the 10 runs for the train, validation and test sets, respectively. The red line describes the variation in Cross-Entropy Loss.

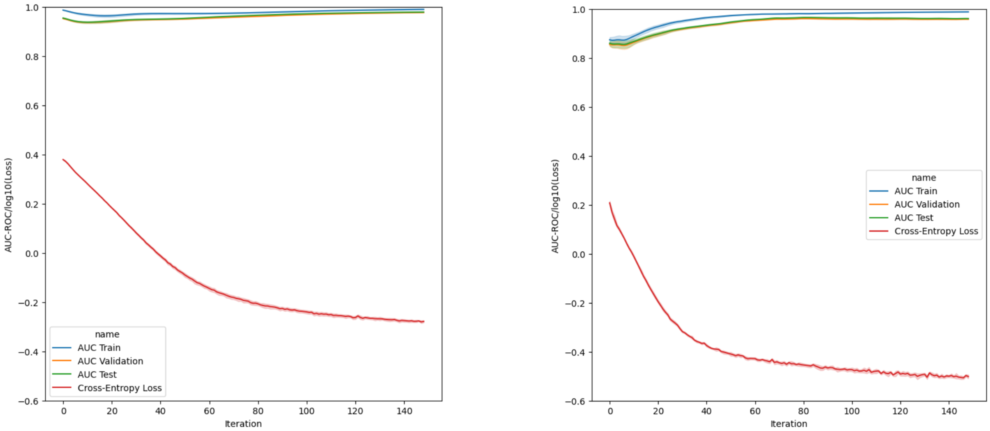
Table 6.2: Summary of the consistency of predictions of all three experiments. The percentages in parenthesis represent the increase of the CM-based values when compared to original KG ones (e.g. increase in mean from original to CM-based).

Disease	DMD		HD		OI	
KG Type	Original	CM-based	Original	CM-based	Original	CM-based
Consist. mean	38.97	**54.1** (+38.82%)	24.27	**48.43** (+99.55%)	11.53	**39.61** (+243.54%)
Consist. median	39.29	**53.57** (+36.35%)	25.29	**52.87** (+109.05%)	10.61	**37.88** (+257.02%)

6.5 Discussion

We extend our results to propose specific research questions (RQs) to guide future studies. Within the context of FAIR and FAIRification, exploring these RQs will enhance understanding of the benefits of FAIR principles for ML applications. Additionally, addressing these RQs will support gathering more data to make the conclusions of our work more generalisable.

Training data. When examining the Cross Entropy Loss curve for OI (Figure 6.6), it becomes evident that the curve of the ML model trained on the CM-based KG is steeper than that of the ML model trained on the original KG. This may suggest



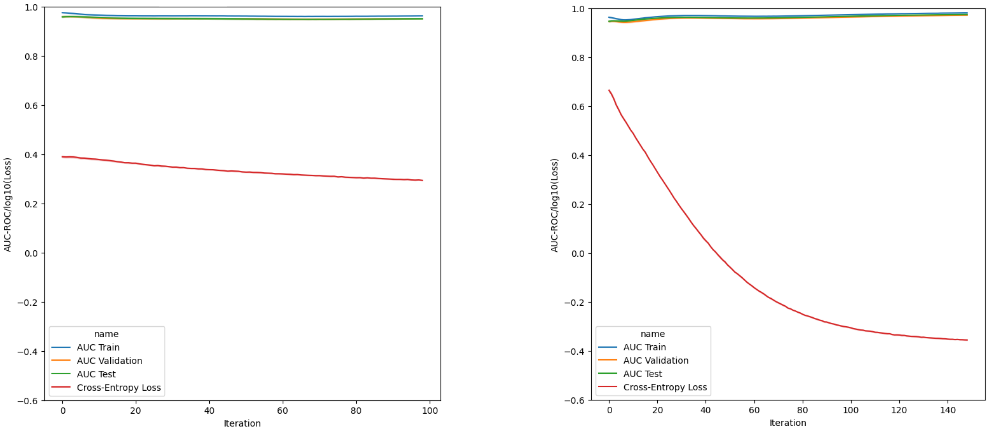
(a) original KG **(b)** CM-based KG

Figure 6.5: Training curves of the GNN models, for **HD**.

that the ML model trained on the CM-based KG "learned more" when compared to the one trained on the original KG. Consequently, future research should explore whether restructuring training data according to well-founded conceptual models can enhance learning in ML models (e.g. GNN models) that initially do not perform well when trained on current data. Thus, a related RQ could be: *"RQ1 - To what extent do better-structured data improve the predictive performance of ML algorithms that underperform on current data?"*

An example of an experiment to address RQ1 could involve identifying cases where data scientists do not manage to further improve the performance of ML models on specific datasets. In such cases, well-founded conceptual models of the datasets' subjects would be designed and used to restructure those datasets and retrain the ML models.

Reproducibility. When comparing the consistency of predictions from ML models trained on the original and CM-based KGs (Figures 6.7, 6.8, and 6.9), it is observed that ML models trained on the latter produce more stable predictions than those trained on the former, as indicated by the average overlap among the 10 lists of predictions generated for each case (Table 6.2). This suggests that training ML models on data structured according to well-founded conceptual models enhances the consistency of predictions, thereby reducing the randomness of the results. Thus, a research question related to this aspect could be: *"RQ2 - To what extent do conceptual model based data contribute to the consistency of predictions of ML models?"*



(a) original KG

(b) CM-based KG

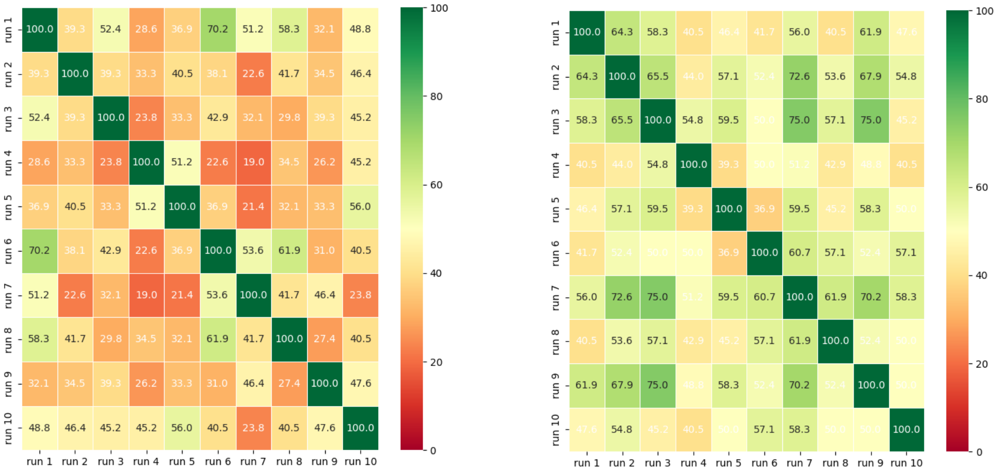
Figure 6.6: Training curves of the GNN models, for **OI**.

To address RQ2, it would be necessary to replicate the experiments conducted in our study in a systematic manner, involving various algorithms and datasets. This approach will ensure that the results are statistically significant and that the conclusions can be generalised and reproduced across different scenarios.

Design of AI systems. During the application of our method, it was observed that conceptual models improved communication, understanding, and task execution among different stakeholders. A pertinent research question arising from this observation is whether data scientists can enhance their performance in feature engineering, ML model selection, and parameter tuning due to a better understanding of the domain data (provided by conceptual modelling tasks). This leads to the final research question: *"RQ3 - To what extent do conceptual models support stakeholders in the design of AI systems?"*

To test RQ3, a controlled experiment could be conducted in which one group of data scientists and developers is tasked with directly designing and implementing an ML pipeline, while another group is required to first create an OntoUML model before proceeding with the design and implementation of the ML pipeline. The results of these two groups would then be compared to evaluate the impact of conceptual modelling on the design and execution of AI systems. Such an experiment could also test whether the conceptual modelling task facilitates communication between data scientists and domain experts.

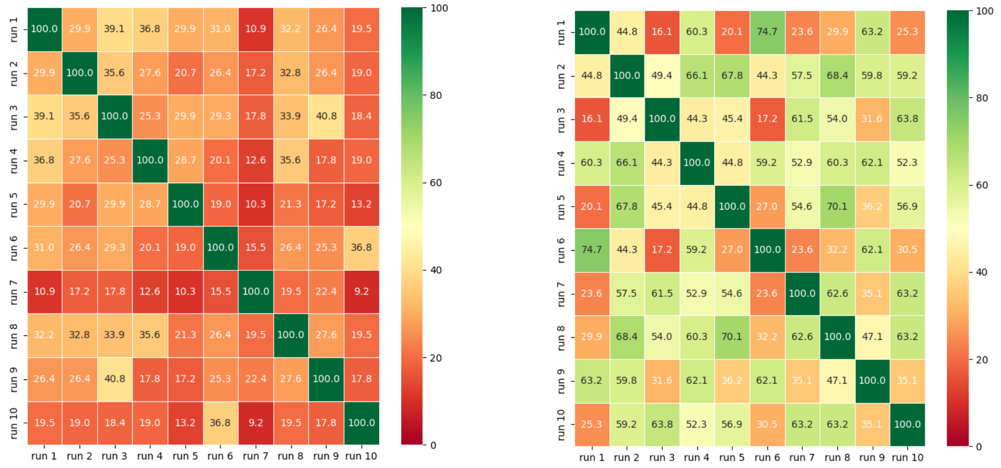
It is expected that the conceptual models produced for the potential experiments



(a) original KG

(b) CM-based KG

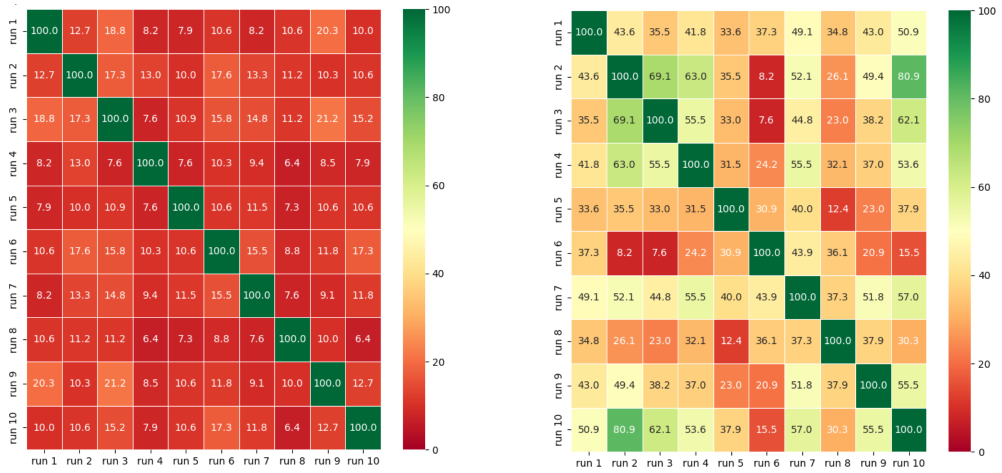
Figure 6.7: Heat map that shows pairwise overlap of predicted drug-symptom pairs of all ten runs in percentages for each case, for **DMD**.



(a) original KG

(b) CM-based KG

Figure 6.8: Pairwise overlap of predicted drug-symptom pairs of all ten runs in percentages for each case, for **HD**.



	run 1	run 2	run 3	run 4	run 5	run 6	run 7	run 8	run 9	run 10
run 1	100.0	12.7	18.8	8.2	7.9	10.6	8.2	10.6	20.3	10.0
run 2	12.7	100.0	17.3	13.0	10.0	17.6	13.3	11.2	10.3	10.6
run 3	18.8	17.3	100.0	7.6	10.9	15.8	14.8	11.2	21.2	15.2
run 4	8.2	13.0	7.6	100.0	7.6	10.3	9.4	6.4	8.5	7.9
run 5	7.9	10.0	10.9	7.6	100.0	10.6	11.5	7.3	10.6	10.6
run 6	10.6	17.6	15.8	10.3	10.6	100.0	15.5	8.8	11.8	17.3
run 7	8.2	13.3	14.8	9.4	11.5	15.5	100.0	7.6	9.1	11.8
run 8	10.6	11.2	11.2	6.4	7.3	8.8	7.6	100.0	10.0	6.4
run 9	20.3	10.3	21.2	8.5	10.6	11.8	9.1	10.0	100.0	12.7
run 10	10.0	10.6	15.2	7.9	10.6	17.3	11.8	6.4	12.7	100.0

(a) Original KG

	run 1	run 2	run 3	run 4	run 5	run 6	run 7	run 8	run 9	run 10
run 1	100.0	43.6	35.5	41.8	33.6	37.3	49.1	34.8	43.0	50.9
run 2	43.6	100.0	69.1	63.0	35.5	8.2	52.1	26.1	49.4	80.9
run 3	35.5	69.1	100.0	55.5	33.0	7.6	44.8	23.0	38.2	62.1
run 4	41.8	63.0	55.5	100.0	31.5	24.2	55.5	32.1	37.0	53.6
run 5	33.6	35.5	33.0	31.5	100.0	30.9	40.0	12.4	23.0	37.9
run 6	37.3	8.2	7.6	24.2	30.9	100.0	43.9	36.1	20.9	15.5
run 7	49.1	52.1	44.8	55.5	40.0	43.9	100.0	37.3	51.8	57.0
run 8	34.8	26.1	23.0	32.1	12.4	36.1	37.3	100.0	37.9	30.3
run 9	43.0	49.4	38.2	37.0	23.0	20.9	51.8	37.9	100.0	55.5
run 10	50.9	80.9	62.1	53.6	37.9	15.5	57.0	30.3	55.5	100.0

(b) CM-based KG

Figure 6.9: Pairwise overlap of predicted drug-symptom pairs of all ten runs in percentages for each case, for **OI**.

described above will be (i) constructed using a well-founded modelling language or ontology as a foundation (e.g. UFO, OntoUML), and (ii) thoroughly validated by domain experts. Finally, it is important to highlight that the RQs described above are formulated from our initial explorations. While some are based on subtle differences in the results (e.g. AUC curves), they remain valuable for further investigation, as they may reveal more significant and impactful differences in other ML applications, particularly those involving large KGs.

6.6 Limitations and future work

Reproducibility and generalisability are the primary limitations of our results. Reproducibility, a well-recognised challenge in ML [244], remains difficult in this context. While our findings demonstrate that CM-based KGs lead to more consistent GNN models, achieving identical results to those presented in Section 6.4 is challenging due, for example, to the inherent randomness in certain components of the *rd-explainer* pipeline. To mitigate this issue, we ran the training and output generation ten times for each case and target disease, averaging the scores to reduce the impact of variability.

Although we synthesised our findings into additional RQs to guide future research, we have not yet tested our findings sufficiently to draw generalisable conclusions. This limitation arises from the fact that our study focused on a single type of ML method, used the same set of data sources within a specific domain, and was constrained by limited computational resources. Therefore, it is crucial to investigate the impact of conceptual models in different domains and with other types of ML methods, as well as with different types of data. For example, while our results may be relevant to graph data and AI methods designed for such data, the application of our method to other data types may not lead to meaningful differences in results.

Finally, an aspect not addressed in this paper but worth considering in future work is the biological basis for the observed behaviours in the experiments. For example, OI has a relatively homogeneous genetic profile, well-documented phenotypic characteristics, and a well-established body of medical knowledge–particularly in terms of phenotypic markers. These factors might simplify the identification of meaningful patterns within the data, potentially contributing to improved prediction accuracy and consistency.

6.7 Related works

The use and impact of conceptual models and ontologies in AI has been a topic of discussion in the literature. Various studies examine how these artefacts can be applied in the field, such as to enhance outcomes and the design of AI systems. At this stage, we view related works as complementary efforts toward a shared ultimate goal: enhancing the understanding and application of ontologies in AI, and vice versa.

Confalonieri & Guizzardi [245] outlined the different roles that ontologies and ontology-based conceptual models can play for neuro-symbolic AI systems from three key perspectives: reference modelling, common sense reasoning, and knowledge refinement and complexity management. Similarly, Maaß & Storey [246] explored the benefits and synergies of integrating conceptual modelling with ML and proposed a framework that uses conceptual modelling to support the design and development of ML solutions.

Lukyanenko *et al.* [247] explored how conceptual modelling can address challenges in extracting insights from large datasets with machine learning. They showed that conceptual modelling aids various ML project phases: defining goals in the business understanding phase, modelling data and identifying quality issues in the data understanding phase, supporting attribute selection and transformation in data preparation, enhancing ML algorithm effectiveness with domain knowledge, improving result interpretability, and documenting process changes during deployment.

6.8 Final Remarks

This work presented an exploratory study to investigate the impact of restructuring knowledge graphs using conceptual models–a step of FAIRification–on the performance of a GNN algorithm. We tested the GNN model's behaviour when applied to both original and CM-based KGs across three different rare diseases targeted for drug repurposing. The initial results provided valuable insights and led to the formulation of additional refined research questions for future investigation, focusing on areas such as supporting the ML design process, improving predictive performance, and enhancing the consistency of predictions.

The most prominent results of this work relate to the consistency of the predictions. We evaluated the prediction lists generated across ten runs of the GNN method on both the original and CM-based KGs for each target disease. This evaluation involved pairwise comparisons measuring the overlap of predictions in each list. In all cases, the prediction lists generated from training the GNN model on the CM-based KGs were more consistent than those generated from the training on the original KGs.

For FAIR and FAIRification, the initial findings herein presented serve as a proof-of-concept of the benefits of applying the FAIR principles to ML applications. Our results demonstrate that FAIR data, structured using well-defined conceptual models, have the potential to enhance the consistency of ML outputs without negatively impacting performance. Additionally, other elements of FAIR further enhance ML design and deployment. For instance, FAIR metadata improve resource discovery (e.g. finding data for training) and simplifies reuse by clearly specifying the conditions under which the data can be reused.

ML and FAIR research are rapidly evolving fields. Our work contributes to this progress by exploring synergies between conceptual models, FAIR principles, and ML. As the next step, we aim to expand our research by applying our method to new domains, diverse data sources, and a broader range of ML approaches for more robust and generalisable results.

Chapter 7

General discussion

This thesis aimed to contribute to facilitating FAIRification and research on FAIR by providing approaches to FAIRification training, goal-based FAIRification planning, and semantic modelling (Figure 7.1). Given the complexity and diversity of research resources in the rare diseases domain, it served as a valuable context for this work in collecting use cases, defining requirements and validating the proposed approaches. This is because the rare diseases field is characterised by heterogeneous and distributed resources, making it a relevant domain to address the challenges of FAIRification.

In the first part of this thesis, challenges of making rare disease resources FAIR were identified and catalogued in the context of the EJP RD. These challenges were then addressed in the subsequent three parts of this work. In part II, formats for FAIR training workshops were described. In part III, GO-Plan, a method to guide FAIRification planning via the identification of objectives was outlined. This method is based on goal-modelling techniques, adapted from the requirements engineering field. Finally, in part IV, the use of foundational ontologies in biomedical research was investigated via a systematic literature mapping aimed at finding experiments that assess the benefits of using such ontologies. Subsequently, in the second chapter of part IV, well-founded conceptual models were applied to the training data of a machine learning-based drug repurposing pipeline in an experimentation setting.

7.1 Main findings

The main findings of this thesis are illustrated in Figure 7.1, and further discussed next. The figure illustrates the contributions of this thesis according to the FAIRification implementation cycle. The contributions of Parts II, III and IV are highlighted in yellow, blue and green, respectively. The contributions from Part I allowed identifying the gaps that motivate the other parts.

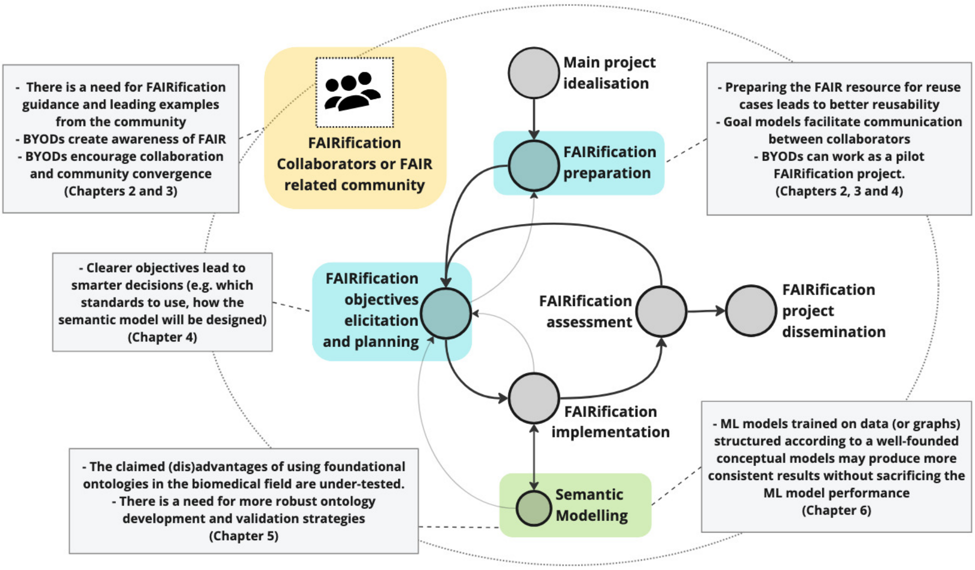

- There is a need for FAIRification guidance and leading examples from the community
- BYODs create awareness of FAIR
- BYODs encourage collaboration and community convergence
(Chapters 2 and 3)
FAIRification Collaborators or FAIR related community
Main project idealisation
FAIRification preparation
- Preparing the FAIR resource for reuse cases leads to better reusability
- Goal models facilitate communication between collaborators
- BYODs can work as a pilot FAIRification project.
(Chapters 2, 3 and 4)
- Clearer objectives lead to smarter decisions (e.g. which standards to use, how the semantic model will be designed)
(Chapter 4)
FAIRification objectives elicitation and planning
FAIRification assessment
FAIRification project dissemination
FAIRification implementation
- ML models trained on data (or graphs) structured according to a well-founded conceptual models may produce more consistent results without sacrificing the ML model performance
(Chapter 6)
- The claimed (dis)advantages of using foundational ontologies in the biomedical field are under-tested.
- There is a need for more robust ontology development and validation strategies
(Chapter 5)
Semantic Modelling

Figure 7.1: Illustration of the contributions of this thesis according to the FAIRification implementation cycle (outlined in Chapter 4). FAIRification collaborators and related community, highlighted in yellow, are supported by contributions from Part II. FAIRification preparation and objectives elicitation and planning, highlighted in light blue, are addressed by contributions from Chatpers 3, 4 and 6. Semantic modelling, highlighted in light green, is supported by the results of Chapters 5 and 6.

7.1.1 Part I: identifying FAIRification challenges

As the FAIR principles gain increased adoption across various fields, it becomes crucial to understand common challenges related to FAIRification and to identify the most significant bottlenecks that should be prioritised. In Chapter 2, challenges faced by partners of the EJP RD project are identified and catalogued by the FAIRification Stewards team, which included the author of this thesis. These challenges include the need for more information and guidance on using FAIR(ification) supporting artefacts such as ontologies (e.g. ORDO), pseudonymisation services, metadata publication software (e.g. FAIR Data Point), semantic models (e.g. the semantic model of the common data elements for rare diseases registration), data exchange standards (e.g. FHIR) and formats (e.g. RDF), querying technologies (e.g. SPARQL), and addressing legal concerns (e.g. capturing informed consent in a machine actionable format).

The FAIRification Stewards team proposed specific solutions to these challenges within the context of the EJP RD. These solutions included providing detailed documentation about FAIR enabling artefacts (e.g. semantic model specifications, ontologies), organising workshops with both technical focuses (e.g. data formats and mapping languages, pseudonymisation and anonymisation tools) and managerial focuses (e.g. practical implementation of registries), as well as training events (e.g. the BYOD-based Rome Summer School), and arranging specialised meetings with experts on particular tools (e.g. experts from the EU RD Platform). Additionally, the team conducted surveys to identify common practices (e.g. a survey on how ERNs annotate disability questionnaires) and organised expert groups to develop solutions to specific issues (e.g. conceptual modelling group).

More broadly, it was observed that organisations aiming to make their resources FAIR have a high demand for expert assistance in several aspects of FAIRification, yet such expertise remains scarce. Furthermore, there is a clear demand for guidance on essential FAIRification steps. For example, there is a need to clarify and detail FAIRification objectives, as this helps inform decisions throughout the process, enabling resources to make more informed choices based on their specific goals. Additionally, having the final goals in mind can increase the motivation of project collaborators, as they gain a clearer understanding of the benefits that the FAIRified resource will bring. Another common challenge identified by the FAIRification Stewards is the need to address conceptual semantic modelling, as resources often seek to align their choices with the most widely used standards and ontologies within their communities.

7.1.2 Part II: building expertise on FAIR

Specialists in FAIR-enabling resources (e.g. ontologies) and FAIRification-supporting tools (e.g. data transformation software) are essential to provide guidance and support increasing the awareness of FAIR. Hence, the BYOD formats described in Chapter 3 are designed to offer initial training to researchers, data scientists, domain experts, and managers. Data-focused BYODs help domain experts in applying the FAIR principles to their data, while management-focused BYODs demonstrate the value of FAIR for project managers. For developers, the software-focused training workshops are a place to ideate the creation or improvement of existing tools, standards, and specifications.

It is suggested that BYODs be adapted to become integral parts of FAIRification projects and workflows. For instance, a BYOD can be organised at the onset of a project, serving as a "FAIRification pilot." This pilot not only creates awareness and motivates the FAIRification collaborators but also helps identify possible project-specific challenges at an early stage. Additionally, BYODs can be used as problem-solving approaches; for example, data-focused BYODs can be employed to design semantic models for the FAIRification project, while software-focused BYODs can be organised to adapt data capture software to generate RDF outputs.

7.1.3 Part III: goal-based FAIRification planning

The identification of FAIRification objectives is an integral part of most FAIRification workflows, as these objectives guide decisions to be made throughout the process. In Chapter 4, the goal-based FAIRification planning method, GO-Plan, is proposed, drawing on previous experience in FAIRification projects and research from the software engineering field. The method is composed of six phases that are further detailed in several steps. The phases are intended to encompass tasks related to project scoping, assessment of the infrastructure and resources to be made FAIR, preparation of stakeholders, identification of reuse cases, goal elicitation and decision making. GO-Plan's validation sessions demonstrated that users view the method as a valuable tool for understanding the FAIRification project, clarifying their roles, making FAIRification more concrete, and reducing communication issues. When applying GO-Plan to a real-world case, it was observed that properly identifying objectives helps establish achievement criteria for principles that lack precision, supports the prioritisation of principles, assists in identifying the best standards to be used in the resulting FAIR resource, and prepares the data for reuse cases. Furthermore, Chapter 4 highlights the need to organise FAIRification projects into different layers and iterations. This can be achieved by segmenting the project into distinct framings. For example, different iterations can focus on various reuse cases or address specific needs of collaborators. For data, iterations can also be structured to focus on manageable data excerpts that are incrementally built upon in subsequent FAIRification implementation cycles.

7.1.4 Part IV: ontology-based semantic modelling for FAIR

Ontologies play a crucial role in biomedical research, as they are used in the standardisation and integration of data sources. Foundational ontologies, in particular, are claimed to enhance the benefits associated with using ontologies and well-founded conceptual models. Foundational ontologies provide a broad, general framework for defining reality concepts (e.g. event, kinds, roles), whereas domain ontologies are more specific and tailored to particular fields or topics. In the systematic literature mapping described in Chapter 5, it was identified that foundational ontologies support the development, analysis, and maintenance of domain ontologies, as well as the integration of data and other standards. Users reported several advantages of using foundational ontologies, as they support improving data understandability and queriability (especially for machines), enhance comprehension and reasoning in domain ontologies, and help the development, maintainability, and interoperability with other ontologies. The most significant disadvantage noted by users of foundational ontologies was their inherent complexity. Additionally, the literature mapping revealed a gap of empirical assessments of the (dis)advantages of using foundational ontologies and a lack of formal methods in building and validating domain ontologies in biomedical research.

Given the lack of experimentation identified in Chapter 5, Chapter 6 explored the impact of conceptual models grounded in foundational ontologies on machine learning (ML). This investigation involved revising and restructuring a knowledge graph (KG) using well-founded conceptual models to train a graph neural network (GNN). The results showed that the ML model trained on the conceptually modelled KG produced predictions that were more consistent (i.e. less random), while the new structure did not affect the performance of the ML model compared to those trained on the original data.

7.2 Strengths and limitations

This thesis has identified and catalogued FAIRification challenges and proposed corresponding solutions to these challenges. A key strength of the results presented in this thesis lies in the novelty of the approaches adopted. To the best of our knowledge, this is the first time that FAIRification challenges specific to rare disease resources have been systematically collected and curated on an international scale. The collection of data from 24 European Reference Networks, each consisting of different institutions in several countries, makes these findings representative for the field. Likewise, the BYODs workshops, which are currently rooted in mature experience from previous events, are the first initiative focused on instructing people while raising awareness on the benefits associated with FAIR.

Novelty is also a strength of the method presented in Chapter 4, and the experiments described in Chapters 5 and 6. GO-Plan is the first published method for FAIRification planning, and it reuses consolidated research results from software engineering research, such as competency questions, goal modelling frameworks and stakeholder analysis, to support FAIRification. Similarly, the experiments conducted to understand the impact of well-founded conceptual models in ML are new to the biomedical domain. Despite the widespread use of ontologies in the field, their claimed advantages and disadvantages are still insufficiently understood, making this the first attempt to assess these aspects.

The limitations of the results presented in this thesis are mainly related to generalisability. The results presented in this thesis have been developed primarily on the basis of experience in rare disease and other biomedical related projects. Other fields may face different challenges and needs that were not present in the biomedical context. In addition to that, for the case of GO-Plan, it should be noted that although the method has been applied in a real-world scenario and validated through two workshops, and is currently used by students of the FAIR Data Engineering course at the University of Twente, it has not yet been used in a large-scale project or in an industrial context. Likewise, the results from the experiments with foundational ontologies are also constrained to their use case. All experiments were conducted using the same method, pipeline and source data, with a particular emphasis on rare diseases. As a result, it remains unclear whether these results can be generalised to other domains or applied to different ML methods. Consequently, further research is needed to explore the broader applicability of both GO-Plan and the results of the experiments with foundational ontologies.

7.3 A reflection on the *status quo*

In this section, I adopt a primarily first-person writing style, as I will be presenting reflections based on my own perspectives. These reflections draw not only from the work and findings presented in this thesis but also from other experiences inherent to most PhD journeys, such as interactions with both junior and senior researchers at conferences, project meetings, and feedback sessions.

While this thesis offers strategies and methods that aim to (i) facilitate the FAIRification process by making it more community-driven, (ii) frame FAIRification through a software engineering lens with a planning method grounded in software engineering principles, and (iii) demonstrate the advantages of FAIR and ontologies through experimental applications, significant challenges remain. First, I have observed that some projects struggle to engage their communities effectively when developing solutions. Second, while the benefits of FAIR are acknowledged, I perceived that some initiatives still tend to overlook key aspects of what a successful FAIRification project requires. Finally, I note that there appears to be hesitancy among some researchers and, at times, even within the community, to fully implement the FAIR principles in their own practices. In the remainder of this section, I delve into these observations further.

FAIR and FAIRification research and solutions must be community-driven.

In Chapter 2, it is emphasised that community-driven solutions are essential for FAIRification, as these solutions should be collaboratively developed by and for the community. Chapter 3 demonstrates the effectiveness of this approach, with BYOD workshops serving as both testing environments for FAIR projects and valuable sources of insights to propel FAIR research.[1] Chapter 4 further reinforces the benefits of this community-driven approach by noting that involving reuse stakeholders directly improves the quality of (meta)data models. Despite these advantages, I have observed that many research groups encounter difficulties in effectively adopting this collaborative approach. This is evidenced by the proliferation of overlapping standards and methods in the state-of-the-art, including redundant ontologies, unused data models, and FAIRification workflows aimed at solving the same or very similar problems. Chapter 5, for instance, has identified several ontologies created with overlapping purposes.

In my view, fostering convergence is essential for achieving a truly interconnected "world of FAIR resources." While a single standard cannot be expected to address every requirement, an excessive proliferation of solutions is also undesirable. Ideally, an ecosystem with a controlled set of continuously evolving standards, well-mapped and interoperable, should be sufficient to meet diverse needs while supporting interoperability and integration. A shift in approach could strengthen convergence: rather than creating new solutions whenever existing ones fall short, we should prioritize adapting, extending, or refining what already exists. Through collaboration in extending and improving current resources, we can avoid creating a "Tower of Babel" of competing standards. Although external influences, such as political factors, may affect the decision to create new standards, these considerations lie beyond the scope of this discussion.

[1] For instance, the generic FAIRification workflow [2] emerged from BYOD experiences, and the FAIR Data Point specifications [71] were initially conceptualized within a BYOD setting.

Given its growing impact on society, FAIRification deserves the rigour of any engineering discipline.

With technology now deeply intertwined with our daily lives, software development has become a massive industry within the global economy. Companies have adopted innovative approaches to software development, treating it with the same rigour as any other engineering discipline. In fact, many software development projects today are even more costly and risky than the construction of physical buildings [248].

This raises an important question: is FAIRification taken as seriously as it should be (e.g. within bioinformatics)? With the growing impact of AI-driven applications, FAIR and ontologised data have become essential. Yet, I have observed that these are often relegated to low-priority status in many projects, with insufficient expertise, budget and time allocated to their implementation.

Here, I reflect on a need for a shift towards a formal "engineerisation" of FAIRification, treating it as an engineering pursuit. The method introduced in Chapter 4 aims to drive this shift by incorporating software engineering principles in itself, thus framing FAIRification as an engineering project in its own right. GO-plan suggests an iterative approach to FAIRification, taking into account organisational resources such as budget, time, and available expertise, which are akin to software development practices.

To support this "engineerisation" of FAIRification, two additional needs must be addressed: first, the need for specialised training to develop more experts in FAIR and ontology design, as discussed in Chapter 3; and second, the need for robust tooling to address various aspects of FAIRification. Professional-grade tools could help mitigate, and potentially automate, some of the complexities inherent in making resources FAIR, such as publishing metadata, transforming raw data into ontologised data, and beyond.

Embracing the new is essential for progress.

While in the first point I advocate for reusing existing solutions to encourage convergence, I also recognise that this is not always the best path. There are situations where emerging challenges require fresh approaches, methodologies and guiding principles, and we must remain open to innovation when current practices fall short.

However, in my experience, I have observed that some researchers are hesitant, or even resistant, to embracing new ideas like the FAIR principles. While the principles inspire enthusiasm in many, especially those poised to benefit from the improvements in making data and resources FAIR, they can also elicit scepticism or reluctance from others. This hesitancy is understandable, as change is never easy. For example, hospitals aiming to adopt FAIR practices for patient data may face substantial adjustments with political, financial, and managerial implications. Such changes may be further complicated by doubts among some stakeholders regarding the tangible benefits that FAIR resources can offer.

Yet, regardless of differing perspectives, the problem remains: data is often not easily findable, access conditions are either unclear or entirely lacking, and data integration and reuse remain challenging. Similarly, the need is clear: for research to continue advancing, resources (especially data) must become easier to reuse. I anticipate, and hope, that resistance to adopting FAIR principles will gradually diminish as more institutions start adopting the principles, thus evidencing their benefits and fostering research.

Given these challenges, I believe it is crucial for organisations embarking on their initial FAIRification efforts to start small, progressively enhancing their systems' alignment with FAIR principles. This gradual approach can help mitigate the initial impact of the transition while fostering both awareness and appreciation of the benefits that FAIR principles bring.

In summary, the FAIR principles hold the potential to transform the digital landscape, but realising this potential requires a thoughtful, collaborative, and adaptive approach. Community-driven efforts, like the ones described in Chapters 2 and 3, can ensure that standards and solutions remain relevant, avoid fragmentation, and support genuine interoperability. As the demand for FAIR and ontologised data grows alongside advances in AI (cf. Chapter 6), it is essential to treat FAIRification and ontology design with the same rigour as any engineering project (cf. Chapters 4 and 5), investing in training, professional-grade tools, and iterative methodologies that account for budget, expertise, and time, such as GO-Plan (Chapter 4).

At the same time, openness to new approaches is equally important. By fostering convergence, adapting established resources where possible, and embracing innovation when necessary, the community can build a sustainable foundation that drives impactful, interoperable research well into the future.

7.4 Future work

The FAIR principles are designed to benefit both researchers and society by accelerating research and discovery, thereby contributing to societal progress. This thesis presents initiatives aimed at facilitating FAIR practices. While these contributions advance the field of FAIR and ontologies research, they also open new opportunities for further research and development.

The BYODs are expected to remain a valuable source of ideas and a space for experimentation that directly contributes to FAIR research. As the FAIR principles mature and are applied across different fields and resources, new types of BYODs are likely to emerge. For example, BYODs focused on creating FAIR vocabularies and ontologies may become necessary in domains where ontologies are not yet widely adopted or developed. Additionally, BYODs could be leveraged to generate their own FAIR training materials, evolving into hands-on workshops used not only to create content but also to design and develop new training programs tailored to specific types of stakeholders. This iterative process would ensure that BYODs continue to meet the evolving needs of the communities they serve, while also expanding their impact in advancing FAIR practices.

For future research on the goal-based FAIRification planning method, a catalogue of FAIRification plans with associated goal models is envisioned. An open-source tool like the Data Stewardship Wizard [249] can be adapted to incorporate the steps of GO-Plan as questions, thereby streamlining the collection of FAIRification objectives. This enhanced tool can be designed to suggest common objectives based on the users' inputs, which would speed up the objectives elicitation process and foster convergence of solutions, while collecting data to augment a FAIRification goal-models catalogue. The proposed tool for identifying FAIRification objectives could be designed to interface with related tools like the Smart Guidance RD Wizard [21], creating a cohesive ecosystem for FAIRification support across various fields.

The experimentation with foundational ontologies and well-founded conceptual models described in Chapter 6 can be extended in several directions. First, to validate the observed impact of conceptual models on ML training data, the experiments should be replicated across different domains and with various ML methods. This approach will help determine whether the benefits of using well-founded conceptual models and ontologies hold true across diverse contexts. Furthermore, research should continue exploring how ontologies influence AI. It is assumed that one of the key advantages of having FAIR resources, particularly data, is their readiness for machine-actionable methods. In this regard, both FAIR data and the underlying ontologised conceptual models can play an important role in informing and enhancing these algorithms. This, in turn, could lead to the development of more effective and transparent AI-based systems. Second, it is aimed to refine the methods discussed in Part IV and further develop it into a goal-based, foundational ontology-driven approach for creating conceptual models for FAIRification. This is intended to facilitate and provide guidance for conceptual modelling for FAIRification.

7.5 Conclusion

This research has contributed to the field of FAIRification, particularly in the context of rare diseases, by identifying and proposing solutions to key challenges through approaches not previously used in research on FAIR. Challenges identified included the lack of expertise and the need for guidance on important FAIRification steps such as objectives identification and semantic modelling. Solutions included training formats, a goal-based method for FAIRification planning, and experiments to understand the impact of foundational ontologies and conceptual models in biomedical applications and in FAIR data. Furthermore, this work has highlighted the importance of community-driven solutions, the need for iterative project planning, and the importance of aligning implementation with FAIRification objectives. While innovative methods such as GO-Plan and the exploration of well-founded conceptual models in AI have shown promise, their generalisability remains limited due to the specific contexts in which they were developed and tested. Nevertheless, these findings provide a strong foundation for future research, particularly in extending these methods to other domains and refining tools and strategies to support the FAIRification process. Finally, the ongoing development of FAIR resources and training initiatives, such as the BYODs, will continue to play a crucial role in promoting the adoption of FAIR principles across different research communities.

Bibliography

[1] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific data. 2016;3:1–9.

[2] Jacobsen A, Kaliyaperumal R, da Silva Santos LOB, Mons B, Schultes E, Roos M, et al. A generic workflow for the data FAIRification process. Data Intelligence. 2020;2(1-2):56–65.

[3] Commission E, Services PE. Cost-benefit analysis for FAIR research data: Cost of not having FAIR research data; 2018.

[4] Bloemers M, Montesanti A. The FAIR funding model: providing a framework for research funders to drive the transition toward FAIR data management and stewardship practices. Data Intelligence. 2020;2(1-2):171–180.

[5] Commission E, for Research DG, Innovation. Turning FAIR into reality – Final report and action plan from the European Commission expert group on FAIR data. Publications Office; 2018.

[6] Kersloot MG, Jacobsen A, Groenen KH, dos Santos Vieira B, Kaliyaperumal R, Abu-Hanna A, et al. De-novo FAIRification via an Electronic Data Capture system by automated transformation of filled electronic Case Report Forms into machine-readable data. Journal of Biomedical Informatics. 2021;122:103897.

[7] Nguengang Wakap S, Lambert DM, Olry A, Rodwell C, Gueydan C, Lanneau V, et al. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. European journal of human genetics. 2020;28(2):165–173.

[8] RD E. The European Joint Programme on Rare Diseases (EJP RD); 2022. https://www.ejprarediseases.org/.

[9] Bernabé CH, Thielemans L, Carta C, et al. Building expertise on FAIR through evolving Bring Your Own Data (BYOD) workshops: Describing the data, software, and management focused approaches and their evolution; 2023.

[10] Guizzardi G. Ontology, ontologies and the “I” of FAIR. Data Intelligence. 2020;.

[11] Horkoff J, Aydemir FB, Cardoso E, et al. Goal-oriented requirements engineering: An extended systematic mapping study. Requirements engineering. 2019;24:133–160.

[12] Van Lamsweerde A. Goal-oriented requirements engineering: A guided tour. In: Proceedings fifth ieee international symposium on requirements engineering. IEEE; 2001. p. 249–262.

[13] Guizzardi G, Baião F, Lopes M, Falbo R. The role of foundational ontologies for domain ontology engineering: An industrial case study in the domain of oil and gas exploration and production. International Journal of Information System Modeling and Design (IJISMD). 2010;.

[14] Guizzardi G, Wagner G, Almeida JPA, Guizzardi RS. Towards ontological foundations for conceptual modeling: The unified foundational ontology (UFO) story. Applied ontology. 2015;.

[15] de Almeida Falbo R. SABiO: Systematic approach for building ontologies. Proc of the 1st Joint Ws on Ontologies in Conceptual Modeling and Inf Systems Engineering. 2014;1201.

[16] Dalpiaz F, Franch X, Horkoff J. iStar 2.0 language guide. arXiv preprint arXiv:160507767. 2016;.

[17] Guizzardi G, Fonseca CM, Benevides AB, et al. Endurant types in ontology-driven conceptual modeling: Towards OntoUML 2.0. In: Conceptual Modeling. ER 2018. vol. 11157. Springer; 2018. p. 136–150.

[18] Guizzardi G, Botti Benevides A, Fonseca CM, et al. UFO: Unified Foundational Ontology. Applied Ontology. 2022;17(1):167–210.

[19] Queralt-Rosinach N, Kaliyaperumal R, Bernabé CH, Long Q, Joosten SA, van der Wijk HJ, et al. Applying the FAIR principles to data in a hospital: challenges and opportunities in a pandemic. medRxiv. 2021;.

[20] Kaliyaperumal R, Wilkinson MD, Moreno PA, Benis N, Cornet R, dos Santos Vieira B, et al. Semantic modelling of common data elements for rare disease registries, and a prototype workflow for their deployment over registry data. Journal of biomedical semantics. 2022;13(1):9.

[21] van Damme P, Moreno PA, Bernabé CH, Ballesteros AC, Le Cornec CM, Vieira BDS, et al. A Resource for Guiding Data Stewards to Make European Rare Disease Patient Registries FAIR. Ubiquity Press. 2023;.

[22] Bernabé C, Keet CM, Khan ZC, Mahlaza Z. A Method to Improve Alignments Between Domain and Foundational Ontologies. In: Formal Ontology in Information Systems. IOS Press; 2023. p. 125–139.

[23] Baldovino S, Moliner AM, Taruscio D, et al. Rare diseases in Europe: From a wide to a local perspective. Israel Medical Association Journal. 2016;18(6).

[24] Hogan Smith K. Review of Rare Diseases Resources: National Organization for Rare Disorders (NORD) Rare Disease Database, NIH Genetic and Rare Diseases Information Center, and Orphanet. Journal of Consumer Health on the Internet. 2017;21(2):216–225.

[25] Saltonstall P, Mike Scott EMD. Toward a focused, multinational, rare disease awareness initiative. In: Rare Diseases: Challenges and Opportunities for Social Entrepreneurs. Routledge; 2017.

[26] Rubinstein YR, Robinson PN, Gahl WA, et al. The case for open science: rare diseases. JAMIA open. 2020;3(3):472–486.

[27] Courbier S, Dimond R, Bros-Facer V. Share and protect our health data: an evidence based approach to rare disease patients' perspectives on data sharing and data protection-quantitative survey and recommendations. Orphanet journal of rare diseases. 2019;14(1):1–15.

[28] European Commission. European Reference Networks; 2023. https://health.ec.europa.eu/european-reference-networks.

[29] Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 2016;3:1–9.

[30] Choudhury A, van Soest J, Nayak S, et al. Personal Health Train on FHIR: A Privacy Preserving Federated Approach for Analyzing FAIR Data in Healthcare. In: Communications in Computer and Information Science. vol. 1240 CCIS; 2020. p. 85–95.

[31] Hallock H, Marshall SE, 't Hoen PAC, et al. Federated Networks for Distributed Analysis of Health Data. Frontiers in Public Health. 2021;9.

[32] Jacobsen A, Kaliyaperumal R, da Silva Santos LOB, et al. A generic workflow for the data FAIRification process. Data Intelligence. 2020;2(1-2):56–65.

[33] Kochev N, Jeliazkova N, Paskaleva V, et al. Your spreadsheets can be FAIR: A tool and FAIRification workflow for the eNanoMapper database. Nanomaterials. 2020;10(10):1908.

[34] Sinaci AA, Núñez-Benjumea FJ, Gencturk M, et al. From raw data to FAIR data: the FAIRification workflow for health research. Methods of information in medicine. 2020;59(S 01):e21–e32.

[35] Groenen KH, Jacobsen A, Kersloot MG, dos Santos Vieira B, van Enckevort E, Kaliyaperumal R, et al. The de novo FAIRification process of a registry for vascular anomalies. Orphanet Journal of Rare Diseases. 2021;.

[36] Kersloot MG, Jacobsen A, Groenen KHJ, et al. De-novo FAIRification via an Electronic Data Capture system by automated transformation of filled electronic Case Report Forms into machine-readable data. Journal of Biomedical Informatics. 2021;122.

[37] EU RD Platform. Set of Common Data Elements; accessed 2023. https://eu-rd-platform.jrc.ec.europa.eu/set-of-common-data-elements_en.

[38] Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: Building an international community of software platform partners. Journal of biomedical informatics. 2019;95:103208.

[39] EvidentIQ. XClinical; 2022. https://xclinical.com/.

[40] RD E. Semantic data model of the set of common data elements for rare disease registration; 2022. https://github.com/ejp-rd-vp/CDE-semantic-model/wiki.

[41] Hub ES. European Rare Disease Registry Infrastructure (ERDRI); 2022. https://eu-rd-platform.jrc.ec.europa.eu/erdri-description_en.

[42] Üstün TB, Chatterji S, Kostanjsek N, et al. Developing the world health organization disability assessment schedule 2.0. Bulletin of the World Health Organization. 2010;88(11).

[43] RD E. EJP RD - European Joint Programme on Rare Diseases - Our Publications; 2022. https://www.ejprarediseases.org/our-publications/.

[44] RD E. ERN Events; 2022. https://ejprd.sharepoint.com/sites/EJPRD-ERN-EVENTS.

[45] Singh J. The portal for rare diseases and orphan drugs. Journal of Pharmacology & Pharmacotherapeutics. 2013;4(2):168.

[46] authors V. International Summer School on Rare Disease Registries and FAIRification of Data; 2022. http://www.ejprarediseases.org/international-summer-school-on-rare-disease-registries-and-fairification-of-data/.

[47] EuRRECa. Data Elements of EuRRECa; 2022. https://eurreca.net/data-elements-2/.

[48] RD E. Hackathon Implementation CDE Semantic Model for ERNs EDC providers; 2022. https://github.com/ejp-rd-vp/EJP-RD-hackathons-workshops/tree/master/EJPRD_Workshop_2020-06_Hackathon_Implementation_CDE_semantic_model_for_ERNs.

[49] RD-CODE. RD Code; 2022. https://www.rd-code.eu/.

[50] RD E. ERN Registries Generic Informed Consent Forms; 2022. https://www.ejprarediseases.org/ern-registries-generic-icf/.

[51] ERICA. The European Rare Disease Research Coordination and Support Action consortium (ERICA); 2022. https://erica-rd.eu/.

[52] RD E. Metadata for EJP rare disease patient registries, biobanks and catalogs; 2022. https://github.com/ejp-rd-vp/resource-metadata-schema.

[53] Dimou A, Vander Sande M, Colpaert P, Verborgh R, Mannens E, Van de Walle R. RML: A generic language for integrated RDF mappings of heterogeneous data. Ldow. 2014;1184.

[54] JRC. SPIDER pseudonymisation tool; 2022. https://eu-rd-platform.jrc.ec.europa.eu/spider/.

[55] Facilitating International Cooperation in Non-Commercial Clinical Trials; 2011. October.

[56] Detlev Gabel TH. GDPR Guide to National Implementation; 2022. https://www.whitecase.com/publications/article/gdpr-guide-national-implementation.

[57] European Union. Regulation 2016/679 of the European parliament and the Council of the European Union. Official Journal of the European Communities. 2016;.

[58] Merrell E, Kelly RM, Kasmier D, et al. Benefits of Realist Ontologies to Systems Engineering. Phillpapers. 2021;.

[59] Hank C, Bishop BW. Measuring FAIR Principles to Inform Fitness for Use. International Journal of Digital Curation. 2018;13(1).

[60] Henning P, Silva LOBd, Pires LF, et al. The FAIRness of data management plans: an assessment of some European DMPs. Revista Eletrônica De Comunicação, Informação e Inovação em Saúde. 2021;.

[61] Schultes E, Magagna B, Hettne KM, et al. Reusable FAIR implementation profiles as accelerators of FAIR convergence. In: Advances in Conceptual Modeling. ER 2020. vol. 12584. Springer; 2020. p. 138–147.

[62] RD E. Introduction to The Advisory Regulatory Ethics Board (AREB); 2022. https://www.ejprarediseases.org/introduction-to-areb/.

[63] RD E. EJP RD General Assembly 2021; 2022. https://www.ejprarediseases.org/ejp-rd-general-assembly-2021/.

[64] RD E. FAIRopoly – FAIRification Guidance for ERN Patient Registries; 2022. https://www.ejprarediseases.org/fairopoly/.

[65] Lusher S, Mons B. Jointly designing a data FAIRPORT; 2014. Lorentz Center at Snellius.

[66] Mark D Wilkinson and Dominique Batista. FAIR Evaluation Services; 2023. https://fairsharing.github.io/FAIR-Evaluator-FrontEnd/.

[67] FAIRSharing Team and Data Readiness Group. FAIRAssist; 2023. https://fairassist.org/.

[68] FAIR Data Team. FAIRifier; 2023. https://github.com/FAIRDataTeam/FAIRifier.

[69] van der Velde KJ, Imhann F, Charbon B, Pang C, van Enckevort D, Slofstra M, et al. MOLGENIS research: advanced bioinformatics data software for non-bioinformaticians. Bioinformatics. 2019;35(6):1076–1078.

[70] W3C. Data Catalog Vocabulary (DCAT) - Version 2. World Wide Web Consortium (W3C); 2019.

[71] da Silva Santos LOB, Burger K, Kaliyaperumal R, Wilkinson MD. FAIR Data Point: A FAIR-oriented approach for metadata publication. Data Intelligence. 2022; p. 1–21.

[72] (W3C) WWWC. Resource Description Framework (RDF); Accessed 2023. `https://www.w3.org/RDF/`.

[73] Creative Commons. Creative Commons Attribution-NonCommercial 4.0 International License; 2023. https://creativecommons.org/licenses/by-nc/4.0/.

[74] EJP-RD. Metadata for EJP rare disease patient registries, biobanks and catalogs; 2023. https://github.com/ejp-rd-vp/resource-metadata-schema.

[75] EJP-RD Consortium. International Summer School on Rare Disease Registries and FAIRification of Data; 2023. https://www.ejprarediseases.org/event/international-summer-school-on-rare-disease-registries-and-fairification-of-data/.

[76] dos Santos Vieira B, Bernabé CH, Zhang S, et al. Towards FAIRification of sensitive and fragmented rare disease patient data: Challenges and solutions in European reference network registries. Orphanet Journal of Rare Diseases. 2022;17:436.

[77] DTL. Successful Hackathon to Make MOLGENIS FAIR; 2016. https://www.dtls.nl/2016/10/30/successful-hackathon-make-molgenis-fair/.

[78] Dutch Techcentre for Life Sciences. First Green BYOD - Yet Another Successful Bring Data Workshop; 2015. https://www.dtls.nl/2015/02/11/first-green-byod-yet-another-successful-bring-data-workshop/.

[79] Dutch Techcentre for Life Sciences. Bring Rett Syndrome Data workshop: a summary; 2016. https://www.dtls.nl/2016/11/10/bring-rett-syndrome-data-workshop-summary/.

[80] Wageningen University & Research. CGN Tomato Collection; 2006. https://www.wur.nl/nl/onderzoek-resultaten/wettelijke-onderzoekstaken/centrum-voor-genetische-bronnen-nederland-1/plant/genebank/cgn-crop-collections/cgn-fruit-vegetables-collection/cgn-tomato-collection.htm.

[81] Dutch Techcentre for Life Sciences. MolData2 implementation study: Facilitating molecular studies of rare diseases; 2018. https://www.dtls.nl/2018/07/06/moldata2-implementation-study-facilitating-molecular-studies-of-rare-diseases/.

[82] Roos M, Gray AJ, Waagmeester A, Thompson M, Kaliyaperumal R, Van Der Horst E, et al. Bring Your Own Data Workshops: A Mechanism to Aid Data Owners to Comply with Linked Data Best Practices. In: SWAT4LS; 2014. p. 1–4.

[83] Gray, Alasdair J G . First FAIRPORT-ELIXIR BYOD workshop; 2014. [Online]. Available at: `https://alasdairgray.github.io/posts/2014-07-11-first-byod-workshop`.

[84] Dutch Techcentre for Life Sciences. About DTL; 2023. https://www.dtls.nl/.

[85] ELIXIR Europe. ELIXIR Europe; 2023. https://elixir-europe.org/.

[86] Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, et al. Towards a knowledge-based human protein atlas. Nature biotechnology. 2010;28(12):1248–1250.

[87] Robert V, Vu D, Amor ABH, van de Wiele N, Brouwer C, Jabas B, et al. MycoBank gearing up for new horizons. IMA fungus. 2013;4:371–379.

[88] Wilkinson MD, McCarthy L, Vandervalk B, Withers D, Kawas E, Samadian S. SADI, SHARE, and the in silico scientific method. In: BMC bioinformatics. vol. 11. BioMed Central; 2010. p. 1–5.

[89] Williams AJ, Harland L, Groth P, Pettifer S, Chichester C, Willighagen EL, et al. Open PHACTS: semantic interoperability for drug discovery. Drug discovery today. 2012;17(21-22):1188–1198.

[90] Slenter DN, Kutmon M, Hanspers K, Riutta A, Windsor J, Nunes N, et al. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. Nucleic acids research. 2018;46(D1):D661–D667.

[91] Gray AJ, Groth P, Loizou A, Askjaer S, Brenninkmeijer C, Burger K, et al. Applying linked data approaches to pharmacology: Architectural decisions and implementation. Semantic Web. 2014;5(2):101–113.

[92] Willighagen EL, Waagmeester A, Spjuth O, Ansell P, Williams AJ, Tkachenko V, et al. The ChEMBL database as linked open data. Journal of cheminformatics. 2013;5(1):1–12.

[93] Lane L, Argoud-Puy G, Britan A, Cusin I, Duek PD, Evalet O, et al. neXtProt: a knowledge platform for human proteins. Nucleic acids research. 2012;40(D1):D76–D83.

[94] Simone Baldovino M, Domenica Taruscio M, Dario Roccatello M. Rare diseases in Europe: from a wide to a local perspective. Isr Med Assoc J. 2016;18:359–63.

[95] RD-Connect. RD-Connect; 2022. Retrieved May 3, 2023, from `https://rd-connect.eu/`.

[96] IRDiRC. Current IRDiRC Recognized Resources; 2023. https://irdirc.org/resources-2/irdirc-recognized-resources/current-irdirc-recognized-resources/.

[97] RD E, editor. International Summer School on Rare Disease Registries and FAIRification of Data. Rome, Italy: Istituto Superiore di Sanità; 2022.

[98] Bernabé C, Sales TP, Schultes E, van Ulzen N, Jacobsen A, da Silva Santos LOB, et al. A goal-oriented method for FAIRification planning. Research Square. 2023;1(Preprint).

[99] Zhang S, Benis N, Cornet R. Automated approach for quality assessment of RDF resources. BMC Medical Informatics and Decision Making. 2023;23(1):1–16.

[100] dos Santos Vieira B, Bernabé CH, Henriques I, Zhang S, Camara AB, García JAR, et al.. Critical steps towards large-scale implementation of the FAIR data principles; 2023. Available from: `https://doi.org/10.5281/zenodo.7867293`.

[101] Bernabé CH, Queralt-Rosinach N, Silva Souza VE, Bonino da Silva Santos LO, Mons B, Jacobsen A, et al. The use of foundational ontologies in biomedical research. Journal of Biomedical Semantics. 2023;14(1):21.

[102] Neuhaus F, Hastings J. Ontology development is consensus creation, not (merely) representation. Applied Ontology. 2022;17(4):495–513.

[103] Hettne, Kristina M and Verhaar, Paul and Companjen, Bas and Kaliyaperumal, Rajaram and Jacobsen, Annika and Burger, Kees and Sesink, Leonie. Bring Your Own Data FAIRification workshop 18 June 2019; 2023. Retrieved from `https://osf.io/8ecg2`.

[104] University of Twente. FAIR Principles and the FAIRification process course; 2021. https://edu.nl/xy96d.

[105] Aktau A, Gambardella A, Hettne K, Ulzen van N. Final outcomes of the taskforce FAIRification as as Service; 2023. Available from: `https://doi.org/10.5281/zenodo.7546767`.

[106] GO FAIR. Making FAIR Metadata; 2023. https://www.go-fair.org/today/making-fair-metadata/.

[107] FAIR G. How to GO FAIR; 2023. https://www.go-fair.org/how-to-go-fair/.

[108] Welter D, Juty N, Rocca-Serra P, Xu F, Henderson D, Gu W, et al. FAIR in action-a flexible framework to guide FAIRification. Scientific data. 2023;10(1):291.

[109] Schultes E. The FAIR hourglass: A framework for FAIR implementation. Fair Connect. 2023;1(1):13–17.

[110] van Lin N, Paliouras G, Vroom E, t Hoen PA, Roos M. How patient organizations can drive FAIR data efforts to facilitate research and health care: A report of the virtual second international meeting on Duchenne data sharing, march 3, 2021. Journal of Neuromuscular Diseases. 2021;8(6):1097–1108.

[111] Engelhardt C, Biernacka K, Coffey A, Cornet R, Danciu A, Demchenko Y, et al. D7.4 How to be FAIR with your data. A teaching and training handbook for higher education institutions. Zenodo; 2022. Available from: `https://doi.org/10.5281/zenodo.6674301`.

[112] Thomer AK, Akmon D, York JJ, Tyler AR, Polasek F, Lafia S, et al. The craft and coordination of data curation: Complicating workflow views of data science. Proceedings of the ACM on Human-Computer Interaction. 2022;6:1–29.

[113] Welter D, Juty N, Rocca-Serra P, Xu F, Henderson D, Gu W, et al. FAIR in action-a flexible framework to guide FAIRification. Scientific Data. 2023;10(1):291.

[114] dos Santos Vieira B, Bernabé CH, Henriques I, Zhang S, Camara AB, García JAR, et al.. Critical steps towards large-scale implementation of the FAIR data principles; 2023. Available from: `https://doi.org/10.5281/zenodo.7867293`.

[115] Pressman RS. Software engineering: A practitioner's approach. 7th ed. McGraw-Hill; 2010.

[116] Freeman RE. Strategic management: A stokcholder approach. Pitman; 1984.

[117] National Information Standards Organization (NISO). Understanding Metadata. Bethesda, MD: NISO Press; 2004. Available from: `https://www.niso.org/publications/understanding-metadata`.

[118] van Damme P, Alarcón Moreno P, Cámara Ballesteros A, Bernabé CH, Le Cornec CMA, Dos Santos Vieira B, et al. A Resource for Guiding Data Stewards to Make European Rare Disease Patient Registries FAIR. Data Science Journal. 2023;.

[119] Queralt-Rosinach N, Kaliyaperumal R, Bernabé CH, et al. Applying the FAIR principles to data in a hospital: challenges and opportunities in a pandemic. Journal of Biomedical Semantics. 2022;.

[120] Davis FD. User acceptance of information systems: the technology acceptance model (TAM). Graduate School of Business, University of Michigan. 1987;Working paper no. 529.

[121] Davis FD, Granić A, Marangunić N. The technology acceptance model 30 years of TAM. Technology. 2023;.

[122] European Joint Programme for Rare Diseases. EJP-RD VP Resource Metadata Schema; 2021. https://github.com/ejp-rd-vp/resource-metadata-schema.

[123] OMG. Business Process Model and Notation (BPMN), Version 2.0; 2011. http://www.omg.org/spec/BPMN/2.0.

[124] Grüninger M, Fox MS. The role of competency questions in enterprise engineering. Benchmarking—Theory and practice. 1995;.

[125] Sansone SA, McQuilton P, Rocca-Serra P, et al. FAIRsharing as a community approach to standards, repositories and policies. Nature Biotechnology. 2019;.

[126] Barcelos PPF, Sales TP, Fumagalli M, et al. A FAIR model catalog for ontology-driven conceptual modeling research. In: Conceptual Modeling. ER 2022. vol. 13607. Springer; 2022. p. 3–17.

[127] Sales TP, Barcelos PPF, Fonseca CM, et al. A FAIR Catalog of Ontology-Driven Conceptual Models; 2023.

[128] Albertoni R, Browning D, Cox S, et al. The W3C Data Catalog Vocabulary, Version 2: Rationale, Design Principles, and Uptake. arXiv preprint arXiv:230308883. 2023;.

[129] Wohlin C, Runeson P, Höst M, Ohlsson MC, Regnell B, Wesslén A. Experimentation in Software Engineering. Springer; 2012.

[130] Kitchenham B, Linkman S, Law D. DESMET: a methodology for evaluating software engineering methods and tools. Computing & Control Engineering Journal. 1997;8(3):120–126.

[131] Bernabé C, Sousa IV, da Silva Santos LOB, Roos M. Insights in FAIRification planning; 2023.

[132] Horkoff J, Yu E. Analyzing goal models: different approaches and how to choose among them. In: Proceedings of the 2011 ACM Symposium on Applied Computing; 2011. p. 675–682.

[133] Kline RB. Principles and practice of structural equation modeling. Guilford publications; 2023.

[134] Bernabé C, Sousa IV, dos Santos Vieira B, Carta C, Jacobsen A, da Silva Santos LOB, et al.. Insights in FAIRification planning; 2024.

[135] Jiang L, Topaloglou T, Mylopoulos J, Borgida A. Goal-oriented conceptual database design. In: 15th IEEE International Requirements Engineering Conference (RE 2007). IEEE; 2007. p. 195–204.

[136] Giorgini P, Rizzi S, Garzetti M. GRAnD: A goal-oriented approach to requirement analysis in data warehouses. Decision Support Systems. 2008;45(1):4–21.

[137] Sothilingam R, Pant V, Shahrin N, Eric S. Towards a Goal-Oriented Modeling Approach for Data Governance. In: PoEM (Forum); 2021. p. 69–77.

[138] Asnar Y, Giorgini P, Mylopoulos J. Goal-driven risk assessment in requirements engineering. Requirements Engineering. 2011;16:101–116.

[139] Van Lamsweerde A, Letier E. Handling obstacles in goal-oriented requirements engineering. IEEE Transactions on software engineering. 2000;26(10):978–1005.

[140] Van Lamsweerde A, Darimont R, Letier E. Managing conflicts in goal-driven requirements engineering. IEEE transactions on Software engineering. 1998;24(11):908–926.

[141] Monfardini GKQ, Salamon JS, Barcellos MP. Use of Competency Questions in Ontology Engineering: A Survey. In: International Conference on Conceptual Modeling. Springer; 2023. p. 45–64.

[142] Splendiani A, Donato M, Drăghici S. Ontologies for bioinformatics. Springer Handbook of Bio-/Neuroinformatics. 2014;.

[143] Keet CM. The use of foundational ontologies in ontology development: an empirical assessment. In: Extended Semantic Web Conference. Springer; 2011. p. 321–335.

[144] Amaral G, Baião F, Guizzardi G. Foundational ontologies, ontology-driven conceptual modeling, and their multiple benefits to data mining. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2021;11(4):e1408.

[145] Consortium GO. The Gene Ontology (GO) database and informatics resource. Nucleic acids research. 2004;32(suppl_1):D258–D261.

[146] Lewis SE. Gene Ontology: looking backwards and forwards. Genome biology. 2005;6(1):1–4.

[147] Bodenreider O, Stevens R. Bio-ontologies: current trends and future directions. Briefings in bioinformatics. 2006;7(3):256–274.

[148] Jackson R, Matentzoglu N, Overton JA, Vita R, Balhoff JP, Buttigieg PL, et al. OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies. Database. 2021;2021.

[149] Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, et al. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic acids research. 2009;.

[150] Stevens R, Wroe C, Lord P, Goble C. Ontologies in bioinformatics. Handbook on ontologies. 2004; p. 635–657.

[151] Hoehndorf R, Schofield PN, Gkoutos GV. The role of ontologies in biological and biomedical research: a functional perspective. Briefings in bioinformatics. 2015;16(6):1069–1080.

[152] Verdonck M, Gailly F, de Cesare S. Comprehending 3D and 4D ontology-driven conceptual models: An empirical study. Information Systems. 2020;93:101568.

[153] Guizzardi G, et al. On ontology, ontologies, conceptualizations, modeling languages, and (meta) models. Frontiers in artificial intelligence and applications. 2007;155:18.

[154] Guarino N. Semantic matching: Formal ontological distinctions for information organization, extraction, and integration. In: International Summer School on Information Extraction. Springer; 1997. p. 139–170.

[155] de Almeida Falbo R, Barcellos MP, Nardi JC, Guizzardi G. Organizing ontology design patterns as ontology pattern languages. In: Extended Semantic Web Conference. Springer; 2013. p. 61–75.

[156] Flügel S, Glauer M, Neuhaus F, Hastings J. When one Logic is Not Enough: Integrating First-order Annotations in OWL Ontologies. arXiv preprint arXiv:221003497. 2022;.

[157] Keet M. Foundational Ontologies; 2020. Available from: `https://eng.libretexts.org/@go/page/6393`.

[158] Smith B, Kumar A, Bittner T. Basic formal ontology for bioinformatics. PhillPapers. 2005;.

[159] Grenon P, Smith B. SNAP and SPAN: Towards dynamic spatial ontology. Spatial cognition and computation. 2004;4(1):69–104.

[160] Iqbal R, Murad MAA, Mustapha A, Sharef NM, et al. An analysis of ontology engineering methodologies: A literature review. Research journal of applied sciences, engineering and technology. 2013;6(16):2993–3000.

[161] Kitchenham B, Charters S. Guidelines for performing systematic literature reviews in software engineering. Information and Software Technology. 2007;.

[162] Gruber TR. A translation approach to portable ontology specifications. Knowledge acquisition. 1993;5(2):199–220.

[163] Studer R, Benjamins VR, Fensel D. Knowledge engineering: Principles and methods. Data & knowledge engineering. 1998;.

[164] Dumontier M, Baker CJ, Baran J, Callahan A, Chepelev L, Cruz-Toledo J, et al. The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery. Journal of biomedical semantics. 2014;.

[165] Beisswanger E, Schulz S, Stenzhorn H, Hahn U. BioTop: An upper domain ontology for the life sciences. Applied Ontology. 2008;3(4):205–212.

[166] Masolo C, Borgo S, Gangemi A, Guarino N, Oltramari A. Ontology Library. WonderWeb Deliverable D18 (ver. 1.0, 31-12-2003); 2003.

[167] Herre H. General Formal Ontology (GFO): A foundational ontology for conceptual modelling. Theory and Applications of Ontology: Computer Applications. 2010;.

[168] Niles I, Pease A. Towards a standard upper ontology. In: Proceedings of the international conference on Formal Ontology in Information Systems; 2001. p. 2–9.

[169] Mizoguchi R. YAMATO: yet another more advanced top-level ontology. In: Proceedings of the sixth Australasian ontology workshop. Citeseer; 2010. p. 1–16.

[170] Vasant D, Chanas L, Malone J, Hanauer M, Olry A, Jupp S, et al. Ordo: an ontology connecting rare disease, epidemiology and genetic data. In: Proceedings of ISMB. vol. 30; 2014. p. 1–4.

[171] Group TB. Journal/Author Name Estimator; 2007. https://jane.biosemantics.org.

[172] for Biotechnology Information (NCBI) NC. PubMed; 2022. https://pubmed.ncbi.nlm.nih.gov.

[173] Elsevier. ScienceDirect; 2022. https://www.sciencedirect.com.

[174] Schulz S, Spackman K, James A, Cocos C, Boeker M. Scalable representations of diseases in biomedical ontologies. In: Journal of Biomedical Semantics. vol. 2. BioMed Central; 2011. p. 1–13.

[175] Pesquita C, Ferreira JD, Couto FM, Silva MJ. The epidemiology ontology: an ontology for the semantic annotation of epidemiological resources. Journal of biomedical semantics. 2014;5(1):1–7.

[176] Hur J, Özgür A, Xiang Z, He Y. Development and application of an interaction network ontology for literature mining of vaccine-associated gene-gene interactions. Journal of biomedical semantics. 2015;6(1):1–10.

[177] Vogt L. Spatio-structural granularity of biological material entities. BMC bioinformatics. 2010;11(1):1–32.

[178] Vogt L. Levels and building blocks—toward a domain granularity framework for the life sciences. Journal of biomedical semantics. 2019;10(1):1–29.

[179] Vogt L, Grobe P, Quast B, Bartolomaeus T. Fiat or bona fide boundary—a matter of granular perspective. PLoS One. 2012;7(12):e48603.

[180] Röhl J, Jansen L. Why functions are not special dispositions: an improved classification of realizables for top-level ontologies. Journal of biomedical semantics. 2014;5(1):1–16.

[181] Jensen M, Cox AP, Chaudhry N, Ng M, Sule D, Duncan W, et al. The neurological disease ontology. Journal of biomedical semantics. 2013;4:1–10.

[182] Brochhausen M, Zheng J, Birtwell D, Williams H, Masci AM, Ellis HJ, et al. OBIB-a novel ontology for biobanking. Journal of biomedical semantics. 2016;.

[183] Tao C, Solbrig HR, Chute CG. CNTRO 2.0: a harmonized semantic web ontology for temporal relation inferencing in clinical narratives. AMIA summits on translational science proceedings. 2011;.

[184] Hoehndorf R, Dumontier M, Oellrich A, Rebholz-Schuhmann D, Schofield PN, Gkoutos GV. Interoperability between biomedical ontologies through relation expansion, upper-level ontologies and automatic reasoning. PloS one. 2011;.

[185] Pisanelli DM, Gangemi A, Battaglia M, Catenacci C. Coping with medical polysemy in the semantic web: the role of ontologies. In: MEDINFO 2004. IOS Press; 2004. p. 416–419.

[186] Brochhausen M, Schneider J, Malone D, Empey PE, Hogan WR, Boyce RD. Towards a foundational representation of potential drug-drug interaction knowledge. In: CEUR workshop proceedings. vol. 1309. NIH Public Access; 2014. p. 16.

[187] Machado CM, Rebholz-Schuhmann D, Freitas AT, Couto FM. The semantic web in translational medicine: current applications and future directions. Briefings in bioinformatics. 2015;.

[188] Martinez-Costa C, Abad-Navarro F. Towards a Semantic Data Harmonization Federated Infrastructure. In: Public Health and Informatics. IOS Press; 2021. p. 38–42.

[189] Masuya H, Makita Y, Kobayashi N, Nishikata K, Yoshida Y, Mochizuki Y, et al. The RIKEN integrated database of mammals. Nucleic Acids Research. 2010;doi:10.1093/nar/gkq1078.

[190] Burek P, Hoehndorf R, Loebe F, Visagie J, Herre H, Kelso J. A top-level ontology of functions and its application in the Open Biomedical Ontologies. Bioinformatics. 2006;22(14):e66–e73.

[191] Keet CM. Transforming semi-structured life science diagrams into meaningful domain ontologies with DiDOn. Journal of biomedical informatics. 2012;45(3):482–494.

[192] Some BMJ, Bordea G, Thiessard F, Schulz S, Diallo G. Design considerations for a knowledge graph: The WATRIMed use case. In: Healthcare of the Future. IOS Press; 2019. p. 59–64.

[193] Boeker M, Jansen L, Grewe N, Röhl J, Schober D, Seddig-Raufie D, et al. Effects of guideline-based training on the quality of formal ontologies: A randomized controlled trial. Plos one. 2013;.

[194] Antoniou G, Harmelen Fv. Web ontology language: Owl. Handbook on ontologies. 2009; p. 91–110.

[195] Rumbaugh J, Jacobson I, Booch G. Unified Modeling Language Reference Manual, The (2nd Edition). London: Pearson Higher Education; 2004.

[196] Kong YM, Dahlke C, Xiang Q, Qian Y, Karp D, Scheuermann RH. Toward an ontology-based framework for clinical research databases. Journal of biomedical informatics. 2011;44(1):48–58.

[197] Barwise J. An introduction to first-order logic. vol. 90. Madison: Elsevier; 1977.

[198] Burek P, Scherf N, Herre H. A pattern-based approach to a cell tracking ontology. Procedia Computer Science. 2019;159:784–793.

[199] Brank J, Grobelnik M, Mladenic D. A survey of ontology evaluation techniques. In: Proceedings of the conference on data mining and data warehouses (SiKDD 2005). Citeseer Ljubljana, Slovenia; 2005. p. 166–170.

[200] Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature biotechnology. 2007;.

[201] Noy NF, McGuinness DL, et al.. Ontology development 101: A guide to creating your first ontology; 2001.

[202] Kassel G. Integration of the DOLCE top-level ontology into the OntoSpec methodology. arXiv preprint cs/0510050. 2005;.

[203] He Y, Xiang Z, Zheng J, Lin Y, Overton JA, Ong E. The eXtensible ontology development (XOD) principles and tool implementation to support ontology interoperability. Journal of biomedical semantics. 2018;9(1):1–10.

[204] Boeker M, Schober D, Raufie D, Grewe N, Röhl J, Jansen L, et al. Teaching Good Biomedical Ontology Design. In: ICBO. Citeseer; 2012. p. 21–25.

[205] Fernández-López M, Gómez-Pérez A, Juristo N. Methontology: from ontological art towards ontological engineering. Proceedings of the Ontological Engineering AAAI-97 Spring Symposium Series. 1997;.

[206] Egana M, Rector A, Stevens R, Antezana E. Applying ontology design patterns in bio-ontologies. In: International Conference on Knowledge Engineering and Knowledge Management. Springer; 2008. p. 7–16.

[207] Gangemi A, Catenacci C, Ciaramita M, Lehmann J. Modelling ontology evaluation and validation. In: European Semantic Web Conference. Springer; 2006. p. 140–154.

[208] Cozzi S, Martinuzzi A, Della Mea V. Ontological modeling of the International Classification of Functioning, Disabilities and Health (ICF): Activities&Participation and Environmental Factors components. BMC medical informatics and decision making. 2021;.

[209] Verdonck M, Gailly F, Pergl R, Guizzardi G, Martins B, Pastor O. Comparing traditional conceptual modeling with ontology-driven conceptual modeling: An empirical study. Information Systems. 2019;.

[210] Hlomani H, Stacey D. Approaches, methods, metrics, measures, and subjectivity in ontology evaluation: A survey. Semantic Web Journal. 2014;.

[211] Peroni S. A simplified agile methodology for ontology development. In: OWL: Experiences and Directions–Reasoner Evaluation: 13th International Workshop, OWLED 2016, and 5th International Workshop, ORE 2016, Bologna, Italy, November 20, 2016, Revised Selected Papers 13. Springer; 2017. p. 55–69.

[212] Fernandes PCB, Guizzardi RS, Guizzardi G. Using goal modeling to capture competency questions in ontology-based systems. Journal of Information and Data Management. 2011;.

[213] Simon J, Dos Santos M, Fielding J, Smith B. Formal ontology for natural language processing and the integration of biomedical databases. International journal of medical informatics. 2006;.

[214] Emeruem C, Keet CM, Dawood ZC, Wang S. BFO Classifier: Aligning domain ontologies to BFO. Research Space. 2022;.

[215] of Twente U. Ontology-Driven Conceptual Modeling with Applications; 2022. https://bit.ly/3q6AOOv.

[216] Nardi JC, de Almeida Falbo R, Almeida JPA. Foundational ontologies for semantic integration in EAI: a systematic literature review. In: Collaborative, Trusted and Privacy-Aware e/m-Services: 12th IFIP WG 6.11 Conference on e-Business, e-Services, and e-Society, I3E 2013, Athens, Greece, April 25-26, 2013. Proceedings 12. Springer; 2013. p. 238–249.

[217] Elmhadhbi L, Karray MH, Archimède B. Toward the use of upper-level ontologies for semantically interoperable systems: An emergency management use case. In: Enterprise Interoperability VIII: Smart Services and Business Impact of Enterprise Interoperability. Springer; 2019. p. 131–140.

[218] Baumgartner N, Retschitzegger W. A survey of upper ontologies for situation awareness. In: Proc. of the 4th IASTED International Conference on Knowledge Sharing and Collaborative Engineering, St. Thomas, US VI; 2006. p. 1–9.

[219] Trojahn C, Vieira R, Schmidt D, Pease A, Guizzardi G. Foundational ontologies meet ontology matching: A survey. Semantic Web. 2022;13(4):685–704.

[220] Partridge C, Mitchell A, Cook A, Sullivan J, West M. A Survey of Top-Level Ontologies - to inform the ontological choices for a Foundation Data Model. CDBB; 2020. Available from: `https://www.repository.cam.ac.uk/handle/1810/313452`.

[221] Brnabic A, Hess LM. Systematic literature review of machine learning methods used in the analysis of real-world data for patient-provider decision making. BMC Medical Informatics and Decision Making. 2021;21(1):54.

[222] Hanson B, Stall S, Cutcher-Gershenfeld J, Vrouwenvelder K, Wirz C, Rao Y, et al. Garbage in, garbage out: mitigating risks and maximizing benefits of AI in research. Nature. 2023;623(7985):28–31.

[223] Jacobsen A, de Miranda Azevedo R, Juty N, Batista D, Coles S, Cornet R, et al.. FAIR principles: interpretations and implementation considerations; 2020.

[224] Guarino N, Guizzardi G, Mylopoulos J. On the philosophical foundations of conceptual models. In: Information modelling and knowledge bases XXXI. vol. 321. IOS Press; 2020. p. 1–15.

[225] Perdomo-Quinteiro P, Wolstencroft K, Roos M, Queralt-Rosinach N. Knowledge Graphs and Explainable AI for Drug Repurposing on Rare Diseases. bioRxiv. 2024;doi:10.1101/2024.10.17.618804.

[226] Ortigossa ES, Gonçalves T, Nonato LG. EXplainable Artificial Intelligence (XAI)–From Theory to Methods and Applications. IEEE Access. 2024;.

[227] Pushpakom S, Iorio F, Eyers PA, Escott KJ, Hopper S, Wells A, et al. Drug repurposing: progress, challenges and recommendations. Nature reviews Drug discovery. 2019;18(1):41–58.

[228] Attia YM, Ewida H, Ahmed MS. Successful stories of drug repurposing for cancer therapy in hepatocellular carcinoma. In: Drug Repurposing in Cancer Therapy. Elsevier; 2020. p. 213–229.

[229] Roessler HI, Knoers NV, van Haelst MM, van Haaften G. Drug repurposing for rare diseases. Trends in pharmacological sciences. 2021;42(4):255–267.

[230] Shefchek KA, Harris NL, Gargano M, Matentzoglu N, Unni D, Brush M, et al. The Monarch Initiative in 2019: an integrative data and analytic platform connecting phenotypes to genotypes across species. Nucleic acids research. 2020;48(D1):D704–D715.

[231] Ursu O, Holmes J, Knockel J, Bologa CG, Yang JJ, Mathias SL, et al. DrugCentral: online drug compendium. Nucleic acids research. 2016; p. gkw993.

[232] Chen X, Ji ZL, Chen YZ. TTD: therapeutic target database. Nucleic acids research. 2002;30(1):412–415.

[233] Hamilton W, Ying Z, Leskovec J. Inductive representation learning on large graphs. Advances in neural information processing systems. 2017;30.

[234] Bernabé C, Zwart R, Perdomo-Quinteiro P, Jacobsen A, Prince Sales T, Queralt-Rosinach N, et al.. Restructuring knowledge graphs with conceptual models - Supplementary material; 2025. Available from: `https://doi.org/10.6084/m9.figshare.28576469`.

[235] Guizzardi G, Fonseca CM, Almeida JPA, Sales TP, Benevides AB, Porello D. Types and taxonomic structures in conceptual modeling: A novel ontological theory and engineering support. Data & Knowledge Engineering. 2021;134:101891.

[236] Janssens ACJ, Martens FK. Reflection on modern methods: Revisiting the area under the ROC Curve. International journal of epidemiology. 2020;49(4):1397–1403.

[237] Hand DJ, Christen P, Kirielle N. F*: an interpretable transformation of the F-measure. Machine Learning. 2021;110(3):451–456.

[238] Mao A, Mohri M, Zhong Y. Cross-entropy loss functions: Theoretical analysis and applications. In: International conference on Machine learning. PMLR; 2023. p. 23803–23828.

[239] Gao Z, Fu G, Ouyang C, Tsutsui S, Liu X, Yang J, et al. edge2vec: Representation learning using edge semantics for biomedical knowledge discovery. BMC bioinformatics. 2019;20:1–15.

[240] Szigyarto CAK, Spitali P. Biomarkers of Duchenne muscular dystrophy: current findings. Degenerative neurological and neuromuscular disease. 2018; p. 1–13.

[241] Walker FO. Huntington's disease. The Lancet. 2007;369(9557):218–228.

[242] Rauch F, Glorieux FH. Osteogenesis imperfecta. The Lancet. 2004;363(9418):1377–1385.

[243] Bergstra J, Bengio Y. Random search for hyper-parameter optimization. Journal of machine learning research. 2012;13(2).

[244] Albertoni R, Colantonio S, Skrzypczyński P, Stefanowski J. Reproducibility of machine learning: Terminology, recommendations and open issues. arXiv preprint arXiv:230212691. 2023;.

[245] Confalonieri R, Guizzardi G. On the multiple roles of ontologies in explanations for neuro-symbolic AI. Neurosymbolic Artificial Intelligence;(Preprint):1–15.

[246] Maass W, Storey VC. Pairing conceptual modeling with machine learning. Data & Knowledge Engineering. 2021;134:101909.

[247] Lukyanenko R, Castellanos A, Parsons J, Chiarini Tremblay M, Storey VC. Using conceptual modeling to support machine learning. In: Information Systems Engineering in Responsible Information Systems: CAiSE Forum 2019, Rome, Italy, June 3–7, 2019, Proceedings 31. Springer; 2019. p. 170–181.

[248] Crosby JP. The Business Manager's Guide to Software Projects. Springer; 2023.

[249] Pergl R, Hooft R, Suchánek M, Knaisl V, Slifka J. "Data Stewardship Wizard": A tool bringing together researchers, data stewards, and data experts around data management planning. Data Science Journal. 2019;18:59–59.

Summary

This thesis offers a comprehensive approach to enhancing FAIR (Findable, Accessible, Interoperable, and Reusable) practices in the biomedical domain, with a particular focus on rare disease research. It addresses the challenges faced by European Reference Networks in making sensitive and fragmented rare disease data FAIR and highlights the crucial role of FAIR data stewards in overcoming various FAIRification challenges, including the need for specialised training, community collaboration, legal compliance, and practical implementation guidance. Building on this, this thesis explores the evolving "Bring Your Own Data" (BYOD) workshops, which are tailored to the needs of different stakeholders, promote collaborative data management practices and identify essential resources for FAIR compliance, creating a dynamic environment for sharing FAIR practices. To further strengthen FAIRification efforts, this work introduces GO-Plan, a goal-oriented planning method that enables communities to set clear FAIRification objectives and align on specific implementation choices. Drawing from software and ontology engineering practices, GO-Plan helps stakeholders develop cohesive FAIRification plans that are adaptable to the unique requirements of diverse biomedical data sources. Additionally, this thesis examines the application of foundational ontologies in biomedical research, highlighting their potential for improving the interoperability of data and other (FAIR) resources. Finally, this research investigates the impact of restructuring knowledge graphs (KGs) with well-founded conceptual models to enhance machine learning (ML) predictions. Findings suggest that KGs restructured with well-founded conceptual models result in more consistent ML predictions without sacrificing model performance, underscoring the potential of conceptual models and FAIR principles in refining ML-driven biomedical applications. Overall, this thesis provides a holistic view of FAIRification in biomedical data, from theoretical foundations to practical applications.

Samenvatting

Dit proefschrift biedt een uitgebreide aanpak voor het verbeteren van FAIR (Findable, Accessible, Interoperable, and Reusable) praktijken in het biomedische domein, met een bijzondere focus op onderzoek naar zeldzame ziekten. Het behandelt de uitdagingen waarmee Europese referentienetwerken worden geconfronteerd bij het FAIR maken van gevoelige en gefragmenteerde gegevens over zeldzame ziekten en benadrukt de cruciale rol van FAIR-gegevensbeheerders bij het overwinnen van verschillende FAIRificatie-uitdagingen, waaronder de behoefte aan gespecialiseerde training, samenwerking binnen de gemeenschap, naleving van wettelijke voorschriften en praktische implementatierichtlijnen. Hierop voortbouwend onderzoekt deze dissertatie de zich ontwikkelende "Bring Your Own Data" (BYOD) workshops, die zijn afgestemd op de behoeften van verschillende belanghebbenden, samenwerkingspraktijken voor gegevensbeheer bevorderen en essentiële bronnen voor FAIR-compliance identificeren, waardoor een dynamische omgeving ontstaat voor het delen van FAIR-praktijken. Om de inspanningen voor FAIR-kwalificatie verder te versterken, wordt in dit werk GO-Plan geïntroduceerd, een doelgerichte planningsmethode die gemeenschappen in staat stelt om duidelijke FAIR-kwalificatiedoelen te stellen en af te stemmen op specifieke implementatiekeuzes. Gebaseerd op software- en ontologie-engineeringpraktijken, helpt GO-Plan belanghebbenden samenhangende FAIRificatieplannen te ontwikkelen die kunnen worden aangepast aan de unieke vereisten van diverse biomedische gegevensbronnen. Daarnaast onderzoekt dit proefschrift de toepassing van fundamentele ontologieën in biomedisch onderzoek, waarbij hun potentieel voor het verbeteren van de interoperabiliteit van gegevens en andere (FAIR) bronnen wordt belicht. Tot slot onderzoekt dit onderzoek de impact van het herstructureren van knowledge graphs (KGs) met goed onderbouwde conceptuele modellen om machine learning (ML) voorspellingen te verbeteren. De bevindingen suggereren dat KG's geherstructureerd met goed onderbouwde conceptuele modellen resulteren in consistentere ML voorspellingen zonder

dat dit ten koste gaat van de prestaties van het model. Dit onderstreept het potentieel van conceptuele modellen en FAIR principes in het verfijnen van ML-gedreven biomedische toepassingen. Over het geheel genomen biedt dit proefschrift een holistische kijk op FAIRification in biomedische gegevens, van theoretische fundamenten tot praktische toepassingen.

Resumo

Esta tese apresenta uma abordagem para aprimorar as práticas FAIR (Findable, Accessible, Interoperable, and Reusable) no domínio biomédico, com foco principal na pesquisa em doenças raras. Ela aborda os desafios enfrentados pelas Redes Europeias de Referência para tornar dados sensíveis e fragmentados sobre doenças raras FAIR, e destaca o papel crucial dos *FAIR data stewards* em lidar com vários desafios da FAIR-ificação, incluindo a necessidade de treinamento especializado, colaboração da comunidade científica, conformidade legal e orientação prática de implementação. Com base nisso, esta tese explora a evolução dos workshops "Bring Your Own Data" (BYOD), que são adaptados às necessidades de diferentes partes interessadas, promovem práticas colaborativas de gerenciamento de dados e identificam recursos essenciais para a dar maior suporte àqueles que desejam mais conformidade com FAIR, criando um ambiente dinâmico para o compartilhamento de práticas e diretrizes. Para fortalecer ainda mais os esforços de FAIRificação, este trabalho apresenta o *GO-Plan*, um método de planejamento orientado por objetivos que permite que as comunidades definam objetivos claros de FAIRificação e se alinhem a escolhas específicas de implementação. Com base nas práticas de engenharia de software e engenharia de ontologias, o GO-Plan ajuda as partes interessadas a desenvolver planos de FAIRificação coesos e adaptáveis aos requisitos exclusivos de diversas fontes de dados biomédicos. Além disso, esta tese examina a aplicação de ontologias de fundamentação na pesquisa biomédica, destacando seu potencial para melhorar a interoperabilidade de dados FAIR. Por fim, esta pesquisa investiga o impacto da reestruturação de grafos de conhecimento (*knowledge graphs* – KGs) com modelos conceituais bem fundamentados para aprimorar métodos de aprendizado de máquina (*machine learning* – ML). Os resultados apresentados sugerem que os KGs reestruturados com modelos conceituais bem fundamentados resultam em predições mais consistentes sem sacrificar o desempenho do modelo de ML, ressaltando o potencial dos modelos conceituais e dos princípios FAIR no refinamento de aplicativos biomédicos orientados por ML. Em geral, esta tese oferece uma visão holística do processo de FAIRificação em dados biomédicos, desde fundamentos teóricos até aplicações práticas.

Acknowledgements
(Agradecimentos)

My PhD journey and the results presented in this thesis would not have been possible without the incredible support of many amazing individuals, both professionally and personally. I am deeply grateful to everyone who has worked with me over the years, as well as to my family and friends who have supported me in countless other ways.

First, I am thankful to my promoters, co-promoters and daily supervisors. Annika, you have been supportive since our first online meeting, when we got to know each other and figured out how to start my PhD amidst the pandemic. You helped me understand the many ways to navigate this journey, you have guided me in shaping my ideas and making them realistic. I am grateful that we were able to work together, and I deeply appreciate your trust in me and my work. I learned a lot from you.

Luiz, nada disso teria sido possível se não tivéssemos nos conhecido no ER 2019 em Salvador. Sou grato por ter me inspirado a seguir essa jornada, e por toda ajuda que me deu desde o início. Suas ideias são inspiradoras, realistas e altruístas.

Marco and Barend, you are both visionaries and your work inspires us all. Barend, your passion for science has motivated me ever since we first met. Marco, you have faith in people and always see the best in them — a quality you have taught me. I am very grateful for your patience, persistence and understanding.

Aos antigos orientadores do mestrado, Vítor, Carla e Renata. O que aprendi com vocês durante o mestrado foi a base para que eu conseguisse continuar esse caminho científico durante o doutorado, muito obrigado. Ao Vítor, agradeço também por ter continuado colaborando e contribuindo com a minha pesquisa, e espero que essa parceria persevere.

I am also grateful to the people in the Biosemantics group. Eleni, thank you for always being there for me. Nuria, you were my companion on the adventure of moving to the Netherlands. I am also grateful to the best students of the Biosemantics group, Daphne and Karolis. Your kindness and joy made me feel welcome in our group and in the Netherlands from the very first day.

I am also thankful for my second family in the Netherlands. Primeiramente, à Bruna e Maggie, por terem me acolhido desde o primeiro momento em que nos conhecemos. Um obrigado especial também à Isadora, por ser um ombro amigo.Iulia, thank you for being so supportive and caring.Agradeço também aos colegas de Twente, em especial Tiago. Obrigado por todo apoio ao meu trabalho, pelos bons conselhos e pela boa vontade em ajudar.

I am also thankful to my EJP RD colleagues, specially to the FAIRification stewards group. I am grateful to have been part of such a impactful project.

Aos amigos que sempre estiveram presentes mesmo na distância, que me deram todo o apoio durante minha adaptação a Holanda, e que sempre fizeram questão de estar comigo nas minhas visitas ao Brasil: Gabi, Shu, Marina, Lucas, Silas, Ana, Gabi T., Gus, Thaís, Pedro e Leo. Muito obrigado! Um agradecimento especial a Miki, que me acompanhou e apoiou em todo esse processo.

Ein herzliches Dankeschön an Sebastian und seine Familie. Vielen Dank für all die Abenteuer, die Unterstützung und die Liebe!

Agradeço a minha família, tia Lila, Lu, Flávia, tio Paulo, Leo e Mazim, aos primos Ana, Vine, Pedro, Isabelly e João Vitor. Obrigado por sempre estarem presentes, por sempre demonstrar entusiasmo pelas minhas conquistas nessa etapa, e por cuidarem de mim e de uns aos outros.

À minha irmã, Maísa. A sua calma e forma humorada de ver o mundo me inspiram. Aos meus pais, Clóvis e Fátima. Vocês são meu exemplo e inspiração. Foi o trabalho duro de vocês que me permitiu ter a oportunidade de trilhar esse doutorado.

Thank you, all. Obrigado a todos.

Curriculum Vitae

César Henrique Bernabé was born in January 1992 in Colatina, Brazil. He attended high school at the Federal Center for Technological Education (CEFET), a federal institution where students receive both high school and technical education. César became a certified construction technician. However, despite his interest in engineering, his true passion had always been informatics.

In 2011, César moved to Vitória to study Computer Science at the Federal University of Espírito Santo (UFES). In 2013, he was selected for the Science Without Borders program, a government-funded initiative that supports high-performing students in a 16-month international exchange. César thoroughly enjoyed his time at the University of Toronto before returning to Vitória in December 2014.

During the second half of his undergraduate studies (2015–2017), César began working as a research intern at the Ontology and Conceptual Modelling Research Group (NEMO), where he continued his work through his master's degree. During his master's studies, he also worked as an ontologist on various projects at the university and other institutions, including Hospital Israelita Albert Einstein (HIAE) in São Paulo. At HIAE, he contributed to a project related to the Brazilian National Health Service (Sistema Único de Saúde - *(viva o)* SUS).

César completed both his bachelor's and master's theses under the supervision of Prof. Dr. Vítor E. S. Souza (UFES) and Dr. Renata S. S. Guizzardi (University of Bolzano). His master's thesis was additionally co-supervised by Prof. Dr. Carla T. L. L. S. Schuenemann (Federal University of Pernambuco - UFPE).

In 2020, César began his PhD research (discussed in this thesis) at Leiden University Medical Center (LUMC), initially working remotely from Brazil due to the COVID-19 pandemic. After being vaccinated, César moved to Amsterdam in 2022 to continue his research and finally meet his colleagues in person. He now lives in Amsterdam and has become an advocate of the FAIR principles.

The figure on this book's cover is an AI-generated illustration of the *Memorial Gratidão* (Gratitude Memorial), a monument created to honour healthcare workers and victims of the COVID-19 pandemic—a period during which most of this thesis was developed. The monument is originally located in Vitória, Brazil, the author's home country, but in the cover image, it is depicted in the Netherlands as a way to symbolically connect both places. Additionally, the monument, which portrays three figures united in harmony, can reflect the need for harmonisation in healthcare (data) to foster research and improve patient diagnosis and treatment (author's reflection).