Improving the interoperability of biomedical research data

Van Damme, P.

Publication date
2023

Chapter 7

General discussion

This thesis aimed to contribute to improving the interoperability of biomedical research data, particularly in the context of rare diseases. First, we examined ways to make data more reusable by providing guidance and harmonization when implementing the FAIR guiding principles (Part I). Second, we studied and developed solutions for challenges around using, developing, and maintaining terminology systems (Part II). In this chapter, we summarize our main findings, address strengths and limitations, discuss the significance and implications of our results, and reflect on future perspectives.

Main findings

Part 1: FAIR Data

What role do the FAIR principles play in guiding and harmonizing data management practices for biomedical research?

The research community widely acknowledges the FAIR principles as the go-to set of guidelines for proper data management to increase the sharing and reuse of data. We set out to investigate the aspects of guidance and harmonization that revolve around the FAIR principles.

Chapter 2 reported on developing a dynamic data management planning questionnaire for data stewards of European rare disease patient registries and addressed 30 challenges that these registries were known to encounter while making their data more FAIR [46]. Moreover, it presented a way to standardize data management practices across multiple patient registries and, thus, improve data interoperability. Chapter 3 presented a case study on implementing the FAIR principles using HL7 FHIR. Most notably, after comparing a deidentified clinical dataset before and after its conversion to FHIR, our study showed that FHIR increased interoperability from 26% to 100% and reusability from 33% to 58%. As a result of these findings, a data standard such as FHIR could simplify some of the challenges and decisions associated with implementing the FAIR principles from the ground up.

Ultimately, the FAIR principles ensure that research communities are on the same page regarding what topics their research data management practices should focus on (e.g., identifier policies, metadata, standard terminology, etc.). Implementers are responsible for choosing the underlying technologies and standards that support the FAIR principles. In this regard, coordination and guidance are key to ensuring that data are FAIR across implementations. The level of interoperability within and among FAIR implementations then depends on the degree of coordination. In other words, once a community decides on what technologies and standards they use, community members should receive guidance that ensures their proper application (Chapter 2). Additionally, if separate implementations should be interoperable, reusing standards and technologies lay the groundwork (Chapter 3). Guiding those who are in charge of implementing the FAIR principles is a two-fold task: (1) given the broad scope of the FAIR principles, guidance on their technical implementation should be tailored to the community and use case for which they are implemented, and (2) the resulting implementation of the principles should be, as far as is feasible, aligned or integrated with existing FAIR implementations and infrastructures.

Part 2: Terminology Systems

What impact do challenges related to the complexity and redundancy of terminology systems have on interoperable data?

Standard terminology is crucial to achieving machine interoperability. Hence, we addressed challenges related to the development, alignment, and quality assurance of terminology systems. In Chapter 4, we described how we transformed a classification for vascular anomalies into an ontology. A patient registry used our ontology to capture machine-readable diagnoses of patients with vascular anomalies. To address the overlap between our ontology and other commonly used biomedical ontologies in the rare disease domain, we added (expert-validated) mappings. In Chapter 5, we assessed the performance of the systems with which we generated these mappings. In this experimental study, we evaluated three state-of-the-art ontology matching systems that obtained precision scores between 0.39-0.54 and recall scores between 0.64-0.96. Accordingly, the evaluated systems were able to retrieve a large number of known mappings but also retrieved a substantial number of mappings whose correctness was unknown. These results suggest that the automatic evaluation of ontology mappings remains challenging.

In Chapter 6, we presented a semi-automatic method for the quality assurance of terminology systems. Our method combined lexical analysis and clustering techniques to identify regularities in the labels of terminology concepts. Based on these lexical regularities, our method proposed logical axioms representing knowledge that should be present in the logical layer of a terminology system. Our method was evaluated using two modules of SNOMED CT and was found to be promising for application in practice. Mainly, it contributed to issues related to incomplete modeling (e.g., detecting that a concept "left ear" does not include a laterality attribute with the value "left").

Overall, our results suggest that the impact of the aforementioned challenges on interoperable data is substantial. Terminology systems are complex to create, use, and maintain, and the related challenges cannot be ignored when one wants to achieve interoperability. Chapters 4 and 5 illustrated that aligning multiple terminology systems is a process that can be supported by (semi-)automatic methods but cannot both be fully automated and reliable. The same goes for quality assurance; our method (Chapter 6) can assist content editors but does not eliminate the manual labor associated with this task.

Strengths and limitations

The work described in this thesis has limitations. First, the scope of the studies we described in Part I is limited to one implementation of a FAIR dataset (Chapter 3) and FAIR infrastructure (Chapter 2). Therefore, other implementations of the FAIR principles, especially those by other research domains, could potentially bear different results. We believe, however, that the examples and lessons our work provides are nonetheless valuable for other domains. Moreover, earlier work based on literature prevalence found that 95% of FAIR implementations were based in the life sciences [206]. Although this proportion is expected to change, our work relates to the majority of current implementations in the biomedical domain.

Another limitation regards the absence of machine learning algorithms in the ontology matching and quality assurance methods described in Chapters 5 and 6, respectively. I.e., algorithms that can be built from training data to predict whether a mapping between concepts or the logical representation of a concept is correct. For example, this could have had the potential to simplify the process of adding mappings to the new ontology we described in Chapter 4. Previous work mentioned that machine learning methods tend to be helpful in matching new terminology systems to a set of already existing systems in the same domain [138].

Furthermore, while real-world use cases drove our studies, we may have failed to include additional aspects or challenges related to implementing the FAIR principles and terminology systems for better interoperability. For example, Thompson et al. have highlighted the importance of mature tools (and lack thereof) that can support humans and machines in making data more FAIR [207]. Although we presented a way to directly support data stewards (Chapter 2), we did not investigate integrating our other methods (Chapters 5 and 6) into tools that aid machine interoperability. We believe that data interoperability is made possible through supporting people as much as solving the technical details.

Our work also has strengths. Terminology systems are essential for machine interoperability and reusable (FAIR) data. Therefore, we consider our strong focus on terminology systems and their associated challenges a significant strength. Matching concepts from overlapping terminology systems, and assuring that the content in these systems is of good quality, will likely remain important in the foreseeable future. Hence, our work adds value to the existing and future scientific work on these topics.

Furthermore, we have a strong focus on real-world use cases from the rare disease domain (Chapters 2, 4, and 5). Given that data and patients in the rare disease domain are geographically scattered, interoperable data is imperative for enabling and advancing research on rare diseases.

Finally, our studies all contain general components, meaning they can be applied to other use cases beyond those described. We consider this another strength of our work. For example, the methodology for developing the dynamic questionnaire in Chapter 2 can be used for other FAIR implementations. A similar idea applies to the design principles and mapping methodology described in Chapter 4. Moreover, our quality assurance method for terminology systems (Chapter 6) can be applied to systems other than SNOMED CT.

Significance and implications

Hughes et al. (2023) identified three main barriers that should be overcome to increase the sharing of FAIR biomedical research data and made suggestions on potential solutions [208]. Table 7.1 summarizes these barriers, their related technical challenges, the suggested solutions, and which chapters in this thesis contribute to them. The following subsections describe how these barriers, challenges, and solutions relate to the work presented in this thesis.

Table 7.1: Barriers, technical challenges, and potential solutions for Findable, Accessible, Interoperable, and Reusable (FAIR) biomedical research data. Based on [208].

Barrier	Technical challenges	Potential solutions	Related chapter
1. No incentives for good metadata, therefore limited data discovery and reuse	- Insufficient metadata - No structured, machine-readable, metadata - No (structured) links to related datasets in publications - No machine-readable version of available data	- Enable access to sufficient resources from research funding - Set up experts in FAIR data to provide guidance and training - Mandate structured metadata and data	Chapter 2
2. Unstandardized methods to access, combine, and generate metadata	- Unstandardized or non-maintained terminology - Terminology systems are not suitable for a specific context - Lack of awareness of existing terminology systems and standards	- Endorse common but extendible metadata standards - Identify concepts across standards that are semantically equivalent - Promote interoperability by providing mappings between commonly used standards - Improve tools that enable (meta)data collection and sharing	Chapter 3 Chapter 5 Chapter 4 Chapter 5, 6
3. Uncoordinated data dissemination efforts		- Provide training and guidance on how to make data FAIR and what this means in specific contexts	Chapter 2

Towards FAIR data and infrastructure

By overcoming the barriers and technical challenges shown in Table 7.1, research data and infrastructure will become more FAIR. In this thesis, we have shown that dynamic data management questionnaires can accommodate guidance on making data and infrastructures FAIR. As demonstrated in Chapter 2, tools such as the Data Stewardship Wizard [49] are readily available for this purpose. Concerning unstandardized (meta)data, our work contributes to solutions related to identifying semantically equivalent concepts between terminology systems and adding such mappings to other systems for better interoperability. Additionally, research communities should focus on coordinating their efforts on sharing FAIR data [208]. Using commonly used standards adapted to specific contexts and use cases is recommended for achieving interoperability. Our work described in Chapter 3 can be helpful here, as a standards framework such as HL7 FHIR allows customization and can be used to implement the FAIR principles, most notably improving interoperability and reusability. Tools should offer features that support users in aligning standard terminology and the quality assurance of terminology systems.

Rare disease and biomedical research data

Challenges for data management in biomedical research include: the lack of sufficiently descriptive metadata; complex or unfeasible ways to access or combine datasets; unstandardized metadata and data; and uncoordinated efforts on data sharing practices [208]. In practice, this means that shared data may not be found by those looking for it (humans or machines), or when they are found, the data are not sufficiently described and can, therefore, not be understood and reused. For rare disease research, one significant consequence of these issues is the difficulty of building cohorts for clinical trials [36]. In an effort to resolve such issues, it has been proposed that research (meta)data about rare diseases should be annotated with concepts from terminology systems [209]. In our work, we found that terminology systems' complexity and redundancy substantially impact interoperable data. Yet, our studies also contribute to using standard terminology from systems aligned with other systems, creating such alignments if they do not exist, and ensuring that the quality of said systems is up to par, which we believe will contribute to better research data interoperability.

Human interoperability

This thesis has mainly focused on improving the machine interoperability of research data. However, the FAIR principles and interoperability go beyond technical challenges [15, 208, 210]. For instance: FAIR guidance depends on humans adhering to what they are guided to do, common standards and terminology systems only work if people agree to use them, and the quality assurance and alignment of terminology systems often still require human intervention. Agreement between people is thus vital for interoperability [211]. Moreover, as we stated in Chapter 1, humans use natural language to be "interoperable". Therefore, we believe that human-readable descriptions of data in natural language should not be replaced with machine-readable descriptions that express meaning using standard terminology. Instead, both should be available for ideal human and machine interoperability. These human-readable descriptions should always reflect the information in the structured data and vice versa. Figure 7.1 depicts an example of how a human-readable summary along with structured data for machines can look like using the HL7 FHIR standard.

```
{
  "resourceType": "Condition",
  "id": "example",
  "text": {
    "status": "generated",
    "div": "<div xmlns=\"http://www.w3.org/1999/xhtml\">Duchenne muscular
dystrophy with an onset age of 5 years</div>"
  },
  "clinicalStatus": {
    "coding": [
      {
        "system": "http://terminology.hl7.org/CodeSystem/condition-clinical",
        "code": "active"
      }
    ]
  },
  "code": {
    "coding": [
      {
        "system": "http://snomed.info/sct",
        "code": "76670001",
        "display": "Duchenne muscular dystrophy"
      }
    ]
  },
     "onsetAge" : {
      "value" : 5,
      "system" : "http://unitsofmeasure.org",
      "code" : "a"
    },
}
```

Figure 7.1: Example of a human-readable summary of machine-readable structured data (in Health Level Seven Fast Healthcare Interoperability Resources [38]). Marked in green is the human-readable narrative. Also note the "display" field revealing that SNOMED CT code 76670001 stands for Duchenne muscular dystrophy.

Future perspectives

To advance science and innovation, research communities will increasingly need to make their data more findable, accessible, interoperable, and reusable. The biomedical research community, and in the context of this work, most notably the rare disease domain, have clear incentives for doing so (e.g., to improve patient care). Although interoperability seems to come after making data findable and accessible, the topics discussed in this thesis around guidance, standards, and terminology systems must be addressed throughout the entire FAIR process. As a first step, research infrastructures should ensure rich (structured) metadata that sufficiently describe what data are about. Machine-readable metadata should use standardized terminology that enables machines to determine what data are described. Thus, standardization can allow machines to identify metadata that describe similar data from various sources.

Institutions, funding agencies, and project consortia should lead future guidance for and harmonization of FAIR implementations, as they are the leading forces in research communities [212]. Individual researchers are not likely to be tasked with these decisions and must, therefore, be made aware of good data management practices (e.g., through training) once decisions regarding the technical implementation of the FAIR principles have been made. Resources such as the dynamic questionnaire we described in Chapter 2 could be a helpful tool for guiding data management personnel. To create awareness of what FAIR data means for a particular implementation, FAIR Implementation Profiles could be used for the dissemination of implementation choices and could additionally support convergence [71]. In addition, using Common Data Elements (common variables and responses that enable consistent and comparable data collection across different research studies or domains) could further increase data interoperability [213, 214].

Recently, large language models such as the Generative Pre-trained Transformer 3.5 (GPT-3.5) and the Bidirectional Encoder Representations from Transformers (BERT) by OpenAI and Google, respectively, have shown the potential to impact structured biomedical data and interoperability [215–218]. These models leverage advanced natural language processing techniques that can assist in converting free-text information into standardized structured data. However, their practical application remains unknown, and issues such as generating false information (hallucination) need to be investigated [219]. Therefore, further research should provide the necessary evidence for confidently applying large language models to the issues described in this thesis.

The ontology matching and quality assurance methods described in Chapters 5 and 6 should be integrated into tools and services to enable their use in practice. This aligns with earlier research highlighting the importance of developing and converging tools and services to support humans and machines in making data FAIR [207, 210].

Finally, further research should continue to look outwards, seeking insights from other domains (best practices, pitfalls, challenges), and applying these learnings to the biomedical domain. For example, earlier publications have showcased FAIR implementations and implementations networks in the fields of ocean observing and chemistry [220, 221]. Sharing implementation choices can improve the coordination between domains and lead to more convergence and interoperability [222].

Conclusion

Interoperable research data can positively influence science. In the domain of rare diseases, interoperability plays a crucial role in facilitating research, including clinical trials, and has a direct impact on the well-being of individuals. In this thesis, we have approached interoperability from the perspective of machines by studying the FAIR guiding principles and terminology systems. Harmonizing FAIR implementations using common standards and guidance is essential for ensuring interoperability within and across domains. Our work contributes to a foundation for developing tools that assist people in overcoming challenges related to building, maintaining, and aligning terminology systems. Having descriptive and structured metadata should be at the start of building research infrastructures. At last, improving the interoperability of our research data requires agreement and collaboration between a broad number of people, from funding agencies to individual researchers. Overcoming technical obstacles related to FAIR data and terminology systems is vital, but they are not the last piece of the interoperability puzzle.