Improving the interoperability of biomedical research data

Van Damme, P.

Publication date
2023

Chapter 1

General introduction

Natural languages, such as English, allow us to express ourselves to our fellow human beings. The beauty of languages is that we can have conversations about almost anything, and even when traveling, we can often communicate what we need to those who speak a language that we do not master or speak poorly. One could say that natural languages allow us to be *interoperable* with other people. While this is trivial for humans, it is not for research data.

Data generation, data use, and data sharing

Data collection is an integral part of the research process [1]. The data could be collected for the first time (i.e., newly generated data), from existing data (i.e., already generated data), or a combination of both. For example, in biomedical research, researchers conducting clinical trials could collect new data from participants' measurements, while researchers conducting retrospective observational studies could collect existing data from Electronic Health Records (EHRs). The research community, including funding agencies, (inter)national policymakers, scientific journals, and research institutions, agree that the sharing of collected research data is essential for science [2–4]. Among the benefits of data sharing is the ability to verify or replicate previously published research, use someone else's data for answering a new research question, or combine datasets to increase sample sizes or the number of included variables [3,5]. However, the proliferation of data sharing and data sharing policies also raises a problem: shared data do not automatically imply data that are usable for others [6,7]. For example, a dataset with blood pressure measurements without information on the measurement technique or patient population is hard to interpret by anyone other than the original creators, let alone by a machine. Machines need structured data and context to process and analyze this information. There has long been a recognition that such metadata are necessary for effective research data management [8]. In addition, data discovery goes hand in hand with good metadata and is an equally important prerequisite for usable data [9]. Sharing (biomedical) data presents several privacy-related challenges, including protecting the privacy of individuals whose data are shared, avoiding unauthorized access and unlawful use, and finding effective ways to deidentify or anonymize data [10–12]. Privacy regulations such as the European Union's General Data Protection Regulation (GDPR) are there to ensure data protection [13]. Although privacy is out of the scope of this thesis, it is an imperative component of data sharing.

All of this fits into the topic of research data reuse, which is when data collected during one research project is reused for another research project [5]. A report from the European Commission, published in 2018, estimated that not having reusable data has a yearly economic impact on the European economy of at least 10.2 billion euros [14]. While making data reusable will require initial and recurrent investments, the same report estimated a net benefit of around 2.6 billion euros, expected to rise over time.

FAIR data equals reusable data

The FAIR Guiding Principles provide guidelines for improving the Findability, Accessibility, Interoperability, and Reusability of research data and infrastructure (Table 1.1) [15]. These

principles have experienced a broad uptake in the research community as a whole and emphasize the need to improve the reusability of data for humans and machines. Findability (how can the data be found?) is about assigning identifiers to (meta)data, ensuring search engines can find the (meta)data, and describing the data with sufficient metadata. Accessibility (how can the data be accessed?) covers the communication protocol via which the (meta)data are accessible and states that metadata should be persistent. Interoperability (how can the data be integrated with other data and applications?) is about serialization formats, standardized terminology, and cross-references. Finally, reusability (can others reuse the data?) concerns licenses, provenance information, and adherence to community standards.

The principles do not dictate the implementation of any particular technology or standard. Implementers are, therefore, free to make their own choices for how to apply the principles to their practice. Many implementations of the FAIR principles have been seen thus far, which has led to different interpretations of what FAIR resembles in practice [16]. Some interpretations have also led to misconceptions about what the FAIR principles stand for (e.g., that FAIR data means open data or that the FAIR principles are a technical standard; neither are true) [17]. Furthermore, due to the high-level nature of the FAIR principles and the subsequent high variety of implementation choices, harmonizing the principles' many possible technical implementations has been challenging [18]. This thesis focuses mainly on the interoperability aspect of FAIR data.

Table 1.1: The 15 FAIR Guiding Principles that provide guidance for making research data and infrastructure more findable, accessible, interoperable, and reusable. Adapted from [15].

FAIR Principle	Description
Findability	
F1	(Meta)data are assigned a globally unique and persistent identifier
F2	Data are described with rich metadata
F3	Metadata clearly and explicitly include the identifier of the data it describes
F4	(Meta)data are registered or indexed in a searchable resource
Accessiblity	
A1	(Meta)data are retrievable by their identifier using a standardized communications protocol
A1.1	The protocol is open, free, and universally implementable
A1.2	The protocol allows for an authentication and authorization procedure, where necessary
A2	Metadata are accessible, even when the data are no longer available
Interoperability	
I1	(Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
I2	(Meta)data use vocabularies that follow FAIR principles
I3	(Meta)data include qualified references to other (meta)data
Reusability	
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
R1.3	(Meta)data meet domain-relevant community standards

Interoperability, the "I" in FAIR

Ide and Pustejovsky defined interoperability as "a measure of the degree to which diverse systems, organizations, and/or individuals are able to work together to achieve a common goal" [19] (p. 2). From an information science perspective on data exchange, there are three levels of interoperability: syntactic, structural, and semantic [20]. Syntactic interoperability is about the format in which data are captured and exchanged. For example, the JavaScript

Object Notation (JSON) is a commonly used data exchange format [21]. Structural interoperability then adds a layer of structure to the data format, the data model. Finally, semantic interoperability adds meaning to the data (model) to allow machines to interpret the data. Figure 1.1 illustrates an example of these three levels, starting from an unstructured sentence in natural language to a structured snippet of data with Schema.org [22] annotations. Adding meaning to data for machines is achieved by using terminology systems (as discussed in the next section), which define meanings of things that exist in the real world, a model of reality [23, 24]. Schema.org, for instance, can describe data on web pages and is used across more than 10 million websites [22].

Terminology systems

Humans can infer implicit knowledge, for example, when reading the sentence "the patient's name is John Doe, born on January 1, 1980". We instantly deduce that "John Doe" is a human name, that they were born in the year 1980 on the first day of the month January, and we understand the notion "patient" as a person who receives some kind of medical care. We, humans, do not need additional semantics beyond our natural languages. Conversely, machines generally need "explicit semantics" that allow for machine interpretation [25]. This is where standard terminology comes into play. Standard terminology is captured in terminology systems, or ontologies. Gruber defined an ontology as "an explicit specification of a conceptualization" and "a specification of a representational vocabulary for a shared domain of discourse — definitions of classes, relations, functions, and other objects" [26] (p. 1). Terminology systems define meanings of things in a structured manner (e.g., "patient" in the example from Figure 1.1). Figure 1.2 depicts a simplified example of two concepts from Schema.org's terminology system. Typically, terminology systems contain content in natural language that is understandable for humans (e.g., labels and descriptions), and logical statements in a computable language that are understandable for machines (e.g., relationships).

Terminology systems are referred to using different terminology, including "thesaurus", "classification", "vocabulary", "nomenclature", "code system", or "ontology" [27]. There are slight differences between these terms, but they are all certain types of terminology systems. In this thesis, we use the terms *terminology system* and *terminologies* as umbrella terms to refer to any of these types of systems. We use the term *ontology* to refer to systems that follow the definition of Gruber (above).

Challenges

Creating, maintaining, and using terminology systems poses various challenges [28–30]. Creating terminology systems means, among other considerations, balancing the granularity and number of concepts one adds to their system. Users of terminology systems often wish for more concepts to choose from [28]. For example, when describing ear infections, one could create the concept "infection" and then add concepts that describe the anatomical parts of the ear and types of infections. Alternatively, one could add concepts for every combination, such as "viral infection of the left ear canal". More concepts increase the system's complexity, making it more challenging to maintain.

Plain text

```
The patient's name is John Doe, born on January 1, 1980.
```

Syntactic interoperability: exchanging the text as JSON enables machines to extract the sentence

```
{
      "text": "The patient's name is John Doe, born on
January 1, 1980."
}
```

Structural interoperability: by applying a data model, the data can be structured, and machines can extract the individual elements

```
{
      "patient": {
            "name": {
                  "givenName": "John",
                  "familyName": "Doe"
            },
            "birthDate": "1980-01-01"
      }
}
```

Semantic interoperability: by adding meaning (semantics) to the elements, a machine can interpret the data and do data integration

```
{
      "@type": [
            "http://schema.org/Patient"
      ],
      "http://schema.org/name": [{
            "http://schema.org/givenName": [{
                  "@value": "John"
            }],
            "http://schema.org/familyName": [{
                  "@value": "Doe"
            }]
      }],
      "http://schema.org/birthDate": [{
            "@type": "http://schema.org/Date",
            "@value": "1980-01-01"
      }]
}
```

Figure 1.1: Three levels of interoperability for machines. From a single sentence in JavaScript Object Notation (JSON) to a structured version with Schema.org [22] annotations.

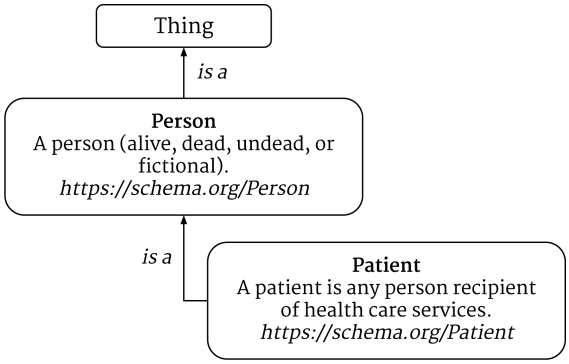

Figure 1.2: Visual representation of concepts in a terminology system, based on Schema.org [22]. Each block represents a *concept*, the bold text represents the *labels*, the text under the labels are *descriptions*, the arrows represent *relationships*, the URLs are *identifiers* (some systems also use codes).

Maintaining large terminology systems requires (semi-)automatic methods to make this a feasible task [30]. For instance, SNOMED Clinical Terms (SNOMED CT), a major clinical terminology system, contains over 358.000 concepts [31]. The quality assurance of such systems is essential, as has been demonstrated by earlier research that found critical errors in SNOMED CT (e.g., one study found that the concept "septic shock" was defined as a "soft tissue infection") [32].

A challenge of using terminology systems is redundancy, both within a terminology system and between systems [28, 29]. Redundancy means that there could be multiple ways to describe the same information using different concepts. For instance, using our earlier example on ear infections, it is possible that a system provides a user with the concept "viral left ear infection" and two separate concepts "viral infection" and "left ear" that a user could use in conjunction. A similar situation can occur between terminology systems, where the same concept is defined by multiple systems [29]. Users then face a choice, which system do you use? Consequently, various users can choose equivalent concepts from different systems to capture the same information. Alignments between terminology systems can solve this issue by exposing equivalent concepts.

The case of rare diseases

Having interoperable data is essential for biomedical research [33]. Biomedical research tends to rely on disparate data sources used by interdisciplinary teams of professionals. Therefore, enabling data integration from different sources and analyzing large data bodies have long been identified as key aspects to advancing biomedical research [34]. The need for interoperable data becomes particularly apparent when examining rare disease research. A study by Ferreira estimated that, on average, a disease is determined to be rare if it has a prevalence of (equal to or lower than) 1 in 2.500 people [35]. While that may seem of little influence, the same study estimated that on a global level, 1 in 16 people suffer from a rare disease, which ac-

cumulates to around 473 million people worldwide. Thus, on a population level, rare diseases are more common than they sound. Due to the dispersed nature of rare disease patients and data, rare diseases present challenges for patients, clinicians, and researchers [36]. Enabling the exchange and integration of data and knowledge can solve some of those issues [37].

Aim and outline of this thesis

This thesis aims to contribute to more interoperable biomedical research data, particularly in the context of rare diseases, by investigating the FAIR principles and terminology systems. We define this research in two parts. First, we aim to explore how to provide technical guidance for implementing the FAIR principles and how such implementations could harmonize their implementation choices for better interoperability (Part I). Second, we aim to investigate the role of terminology systems for FAIR data and address challenges related to development, cross-terminology use, and quality assurance (Part II). Chapters 2, 4, and 5 of this thesis use examples from or are directly related to the field of rare disease research. Following are our research questions:

1. What role do the FAIR principles play in guiding and harmonizing data management practices for biomedical research?
2. What impact do challenges related to the complexity and redundancy of terminology systems have on interoperable data?

Part I: FAIR Data

The FAIR principles set out to improve the reusability of research data and infrastructure [15]. Part I of this thesis explores the FAIR principles from an implementation and harmonization perspective in the context of biomedical research data. Chapter 2 presents a data management planning tool to guide data stewards of European rare disease patient registries. As a dynamic questionnaire, this tool addresses issues that patient registries encounter when implementing the FAIR principles. Chapter 3 focuses on how Health Level Seven Fast Healthcare Interoperability Resources (HL7 FHIR) [38], a data exchange standard for healthcare, can be used to implement the FAIR principles. We present a case study on a real-world deidentified critical care dataset and propose a set of implementation choices, in an effort to contribute to more aligned FAIR implementations.

Part II: Terminology Systems

As pointed out previously, the use of terminology systems plays a vital role in achieving data interoperability. Part II of this thesis focuses on the development, alignment, and quality assurance of terminology systems. Chapter 4 describes the result of developing an ontology for vascular anomalies, which should enable the structured registration of such diagnoses. In addition, it reflects on design decisions that were made while building the ontology. Chapter 5 presents an experimental study to assess the performance of three automated matching systems that can produce alignments between ontologies. It uses examples from the rare

disease research domain and aims to contribute to the interoperability aspect of FAIR data. Chapter 6 presents a method for the semi-automatic quality assurance of terminology systems. The method identifies lexical regularities in the natural language content of terminology systems and proposes logical axioms (statements) that should be present in the logical layer of the system. This method was applied to two modules of SNOMED CT.

Finally, Chapter 7 provides a general discussion of the work presented in this thesis and discusses the main findings and future perspectives.