Ontological representation of FAIR principles: A blueprint for FAIRer data sources

Anna Bernasconi[1,⋆][0000−0001−8016−5750] Alberto García S.[2,⋆][0000−0001−5910−4363], Giancarlo Guizzardi[3][0000−0002−3452−553X], Luiz Olavo Bonino da Silva Santos[3,5][0000−0002−1164−1351], and Veda C. Storey[4][0000−0002−8735−1553]

[1] Politecnico di Milano, Milan, Italy anna.bernasconi@polimi.it
[2] Universitat Politècnica de València, Valencia, Spain algarsi3@pros.upv.es
[3] University of Twente, Enschede, The Netherlands g.guizzardi@utwente.nl,l.o.boninodasilvasantos@utwente.nl
[4] Georgia State University, Atlanta, Georgia, USA vstorey@gsu.edu
[5] Leiden University Medical Center, Leiden, The Netherlands

Abstract. Guidelines to improve the Findability, Accessibility, Interoperability, and Reuse of datasets, known as FAIR principles, were introduced in 2016 to enable machines to perform automatic actions on a variety of digital objects, including datasets. Since then, the principles have been widely adopted by data creators and users worldwide with the 'FAIR' acronym becoming a common part of the vocabulary of data scientists. However, there is still some controversy on how datasets should be interpreted since not all datasets that are claimed to be FAIR, necessarily follow the principles. In this research, we propose the *OntoUML FAIR Principles Schema*, as an ontological representation of FAIR principles for data practitioners. The work is based on OntoUML, an ontologically well-founded language for Ontology-driven Conceptual Modeling. OntoUML is a proxy for ontological analysis that has proven effective in supporting the explanation of complex domains. Our schema aims to disentangle the intricacies of the FAIR principles' definition, by resolving aspects that are ambiguous, under-specified, recursively-specified, or implicit. The schema can be considered as a blueprint, or a template to follow when the FAIR classification strategy of a dataset must be designed. To demonstrate the usefulness of the schema, we present a practical example based on genomic data and discuss how the results provided by the OntoUML FAIR Principles Schema contribute to existing data guidelines.

Keywords: FAIR data · OntoUML FAIR Principles Schema · FAIRness guidance · Ontological Modeling Language · OntoUML

⋆ A.B. and A.G.S. should be regarded as Joint First Authors.

1 Introduction

Since the publication of the seminal paper in early 2016 [33], the FAIR principles gained significant attention. The principles have quickly been embraced by industry and academic communities, leading to several initiatives aiming at developing FAIR-compliant implementations. The implementation efforts can normally be classified into FAIRness assessment [34,1,6], FAIR tooling [30] and FAIR service support [28,20].

However, these initiatives quickly faced challenges to consistently interpret the principles in enough detail that could derive proper implementations. A possible consequence of inconsistent interpretations of the principles is the emergence of potentially incompatible implementations, defeating the original purpose of the FAIR principles. The difficulty in consistent interpretation can be traced to two aspects: (i) by design, principles do not provide specific implementation definitions; and (ii) the original FAIR paper did not explain in detail the intentions behind the principles and related consequences. Almost four years later, a subset of the original FAIR paper authors, together with other collaborators attempted to provide further explanations for the intended interpretations of the FAIR principles and implementation considerations related to each principle and sub-principle [19].

Once someone intends to adopt the FAIR principles to "make my data FAIR", three main questions need to be answered: (i) to what extent does my resource (e.g., dataset) currently follow the FAIR principles, i.e., what is its current FAIRness level?; (ii) what is the intended FAIRness level that I want it to reach?; and (iii) how can I improve from the current-level to the intended-level FAIRness? To answer these questions, one must hold a good understanding of the principles. Ontological models, traditionally employed to provide clear and precise explanations of a domain and enforce its shared understanding among stakeholders, are particularly suitable for representing the complex world of FAIR principles.

The objective of this research, therefore, is to propose the *OntoUML FAIR Principles Schema* resulting from an ontological analysis of the FAIR principles. OntoUML is an ontologically well-founded language for Ontology-driven Conceptual Modeling, which was built as a UML extension based on the Unified Foundational Ontology (UFO) [14]. This means that the modeling primitives of the language reflect the ontological distinctions put forth by the underlying foundational ontology. In other words, the modeling patterns constituting the language reflect the axiomatic micro-theories in UFO [27]. As a proxy for ontological analysis, OntoUML has proven to be a very effective support for the explanation of complex domains [16,9,3], because it can improve the understandability of technical concepts over traditional conceptual models [32,8].

Information integration and interoperability are important for the Information Systems domain. These aspects can be facilitated by applying the FAIR principles to both the data handled by information systems and the systems themselves. We aim to demonstrate that a clear and precise description leads to a more consistent interpretation of the FAIR principles and can contribute to

the Information Systems domain by facilitating the development of more interoperable information ecosystems.

The contribution of our work is to show how using a foundational ontology-based model to represent the FAIR principles can provide the following benefits: (a) Explicit representation of a particular shared interpretation of the principles in a concrete artifact; (b) Controlled vocabulary for use in semantic annotations of (meta)data entities; (c) Rationale for deriving FAIR evaluation metrics; (d) Prescriptive guidelines based on the metrics that operationalize the more abstract guiding principles. Moreover, the ontological schema also facilitates the use of its concepts and relations to semantically annotate metadata and data to make explicit, to machines and humans, their semantic commitments with the interpretation of the FAIR principles that our proposed schema represents.

The remainder of the paper is organized as follows. Section 2 provides an overview of the FAIR principles. Section 3 briefly describes the modeling language applied, OntoUML, the modeling method, and the resulting Schema (i.e., the main result of this paper addressing benefit (a) above). Section 4 shows one application in the genomics domain to illustrate how the schema can be used for semantic annotation (benefit (b)) and for deriving prescriptive operationalization guidelines (benefit (d)). Section 5 discusses the implications and Section 6 concludes the paper.

2 The FAIR Guiding Principles

FAIR Principles were first proposed by Wilkinson et al. [33] and further discussed and analyzed by the GO-FAIR initiative (`https://www.go-fair.org/fair-principles/`). The four FAIR principles are divided into the following sub-principles.

Findability. The utility of a dataset depends, to a large extent, on how easily its potential users can find it. The FAIR principles consider both humans and computers as potential users. Therefore, the means to uniquely identify a given digital object and the provision of rich enough metadata - so that potential (re)users can discover it - are the main targets of the findability sub-principles described below.

F1. (meta)data are assigned a globally unique and persistent identifier. Both metadata and data should be uniquely identified by persistent identifiers (e.g., a globally unique and persistent URI). To ensure uniqueness, once an identifier has been associated with a (meta)data, the same identifier should not relate to any other object. The persistence aspect relates to the identifier being associated with the same object over a period of time. F1 is one of the most relevant FAIR principles, because several others are built upon unique identifiers. Commonly, data repositories automatically assign globally unique identifiers for their hosted datasets. However, it is not always the case for the metadata records.

F2. data are described with rich metadata (defined by R1 below). Although the distinction between data and metadata is arbitrary, this principle attributes to metadata the specific role of describing other data with, for

instance, the types of descriptors defined in the R1 sub-principles. The metadata should be "generous and extensive". The richer the metadata, the higher the chance that potential users find data based on the information provided in the metadata. Several types of metadata exist, including data about how a dataset was processed (e.g., the assembly used in a sequencing process); the context surrounding its acquisition (e.g., the protocol for obtaining a biological sample); device measurements (e.g., quality data for the devices used to extract a biological sample); or domain-specific information (e.g., the genes or proteins considered in a sequencing procedure).

F3. metadata clearly and explicitly include the identifier of the data it describes. In some approaches or technologies metadata and data are stored in the same location (e.g., readme-files in the same folder as the described file) or even in the same file (e.g., EXIF metadata in image files). However, we cannot rely on file proximity to establish the connection between the metadata record and the object it describes. When someone discovers a given dataset through its metadata, the metadata record should explicitly contain the identifier of the dataset. Since a metadata record may contain a number of different identifiers (e.g., identifiers of other concepts and relations), the data identifier should be indicated with a relation/predicate that clearly communicates their connection (e.g., the isMetadataOf relation in our proposed schema).

F4. (meta)data are registered or indexed in a searchable resource. Although rich metadata increases datasets' findability, it is insufficient to ensure it. If the existence of a dataset is unknown, users will not be able to use it, regardless of its metadata. Datasets and/or their corresponding metadata should be indexed in searchable engines. Principles F1, F2, and F3 establish the core requirements for findability and principle F4 indicates the finding mechanism.

Accessibility. When a user finds a (meta)data and, consequently its identifier, there should be a mechanism to access the (meta)data. Moreover, in many situations, data will cease to exist. Even then, it is relevant to keep their metadata accessible so the existence and characteristics of the data can still be known.

A1. (meta)data are retrievable by their identifier using a standardized communication protocol. Datasets are retrieved with the support of a communication protocol. Among the several available ones, some are private, offer limited implementation capabilities, or are poorly documented. To address these cases, this FAIR principle requires the standardized communication protocol used to retrieve the (meta)data from its identifier to have the following properties: 1) the communication protocol should be open, free, and universally implementable; 2) the communication protocol should provide authentication and authorization procedures when required. FAIR does not require the data or metadata to be open or free; rather, it requires that the descriptions of the mechanism to retrieve the data and/or metadata be open and free (A1.1).

A2. metadata are accessible, even when the data are no longer available. The metadata associated with a dataset is a valuable resource *per se*. Data tend to become inaccessible over time for a variety of reasons, e.g., unsustainable

maintenance costs. If a dataset becomes inaccessible, its corresponding metadata should remain accessible. In practice, metadata is much easier and cheaper to maintain accessible.

Interoperability. This principle considers the ability of (meta)data from one source to be connected in workflows with (meta)data from other sources for different purposes, such as analysis, storage, or processing.

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation. For data to be exchangeable, it should be in the same format, or appear in formats that can be parsed and interpreted by its corresponding parties. Therefore, this principle determines minimal requirements for the language used to represent the (meta)data.

I2. (meta)data use vocabularies that follow FAIR principles. Another aspect to improve interoperability is the use of (commonly) employed vocabularies, ontologies, thesauri or data models. However, these vocabularies should also be findable, accessible, interoperable and reusable to a certain degree.

I3. (meta)data include qualified references to other (meta)data. For links to be meaningful, references between entities must have a clear and informative semantics ('is regulator of' is better than 'is associated with'). This holds for both metadata elements and entire datasets. Datasets should be cited and the scientific links between them described.

Reusability. For optimizing the reuse, metadata and data should be well-described so their use in different settings can be assessed.

R1. meta(data) are richly described with a plurality of accurate and relevant attributes. If deemed useful in a particular context, data should be reused. In order to support the potential (re)user to assess whether a particular data is relevant, rich and relevant metadata information should be provided. The R1 sub-principles provide 3 main categories of relevant metadata information to foster reuse as well as compose the rich metadata expected by principle F2. A non-exhaustive list of exemplar metadata properties is provided at (`https://www.go-fair.org/fair-principles/r1-metadata-richly-described-plurality-accurate-relevant-attributes/`).

R1.1. (meta)data are released with a clear and accessible data usage license. This sub-principle refers to the usage rights attached to a dataset, from which one can define the possible legal interoperability (licensing) of the data. Moreover, by defining reuse conditions, which include accessing the (meta)data, the license determines when the authentication and authorization procedures of the sub-principle A1.2. are required.

R1.2. (meta)data are associated with detailed provenance. This principle refers to the information about how the (meta)data came about, e.g., its origin and history, including who oversaw its generation and how it did so.

R1.3. (meta)data meet domain-relevant community standards. When (even minimal) information standards exist for a community, these should be employed to make similar (meta)data easier to be used together.

3 Ontological Modeling

3.1 OntoUML

The meta-model of the OntoUML language complies with the ontological distinctions and axiomatization of the well-grounded Unified Foundational Ontology (UFO [17]). Only the main concepts are presented here. The complete presentation of the language, the philosophical justifications, formal characterizations, and primitives are available in [14].

OntoUML includes the subcategory of *endurants*, i.e., entities that have essential and accidental properties and, hence, can change over time. Endurant types include *Kinds*, the fundamental types of objects that exist in a domain. Objects classified by a kind could not possibly exist without being of that specific kind. All objects necessarily belong to exactly one kind and cannot change kinds. There can be other static subdivisions of a kind, termed *Subkinds*. Object kinds and subkinds represent essential properties of objects (*rigid* types). There are, however, types that represent contingent or accidental properties of objects (*anti-rigid* types). These include *Phases* (properties that are intrinsic to entities: 'being a puppy' is being a dog in a particular developmental phase) and *Roles* (properties of entities within a relational context: 'being a husband' captures a cluster of contingent relational properties of a man participating in a marriage). Kinds, Subkinds, Phases, and Roles are categories of object *Sortals* (a type that provides a uniform principle of identity, persistence, and individuation for its instances).

Relators (such as enrollments, mandates, affiliations) represent clusters of relational properties that are kept together by a nexus. Relations (as classes of n-tuples) can be completely derived from relators [12]. Relators are existentially dependent entities that bind together entities (their relata) by the *mediation* relations, which is a particular type of *existential dependence* relation (A being existentially dependent on B means that B has to exist in all situations where A exists). Besides existential dependence, OntoUML countenances the relation of *external dependence*: an object A is externally dependent on an object B iff A is existentially dependent on B and B and A are mereologically disjoint (neither A is part of B nor B is part of A and they do not share any common part). Objects typically participate in relationships (relators) playing certain "roles". We call *RoleMixins* those role-like types that classify entities of multiple kinds.

Types that represent properties shared by entities of multiple kinds are called *Non-Sortals*. *Categories* are non-sortals that represent necessary properties shared by entities of multiple kinds.

Objects have parts (called components) that play different functional roles with respect to the whole. *Collectives* are entities that have a uniform structure, i.e., whose parts play the same role with respect to the whole.

Besides relator, another type of dependent endurant is a *mode*: an endurant that is existentially dependent on (inheres in) a singular individual.

3.2 OntoUML FAIR Principles Schema

Operationally, we first considered the FAIR principles as a whole, capturing the general spirit they convey. Then, we considered one sub-principle at a time, evaluating which OntoUML entity stereotypes are needed for representing the involved concepts and then connecting them with relationship stereotypes or generalization/compositions. In the following, we describe the OntoUML FAIR Principles Schema, representing the template of a world that more precisely instantiates FAIR choices within the context of the creation of a scientific dataset.

Findability and Interoperability. The excerpt of the schema related to these two highly interlinked principles is shown in Fig. 1. Data is a *collective* of Data Items. One should notice that Metadata is Data, namely, data that *refers to* (*is externally dependent of*) Data. Metadata describes, puts into context, and informs the provenance of data. The recursive chain of reference here stops at what is termed Ground Data, i.e., data that cannot serve as metadata to other data. Metadata does not have to have metadata itself – otherwise, we would incur in a vicious regression. However, ground data must be described by metadata (see **F2**). Data can play the role of Searchable Data (i.e., data with a metadata description) when it benefits from a Registration/Indexing relator, within a Searchable Resource (see **F4**). Data Items are, in turn, composed of Attributes. Since metadata is data, Metadata Items are those data items that compose metadata resources and, analogously, Metadata Item Attributes are those attributes that compose metadata items. Note that (meta)data and their (meta)data item subparts (termed here simply Data Entities) must conform to a Data Model, expressed by a Representation Language – which could be different for each of these (see **I1**). Data Models and Representation Languages are Resources. Any resource can be considered as a Community Standard. The acceptance of a resource as a community standard is given by the Community Consensus of a given Community (a collective belief of a community, thus, modeled here as a *mode*), see **I1**. For example, Data could be instantiated with a textual genomic file, following an XML schema. Each item could be a genomic region following the simple schema <chromosome, start_coordinate, end_coordinate>. The linked Metadata could be another textual file – each referring to one data file – composed of <key,value> pairs items. Every attribute is composed of two essential and inseparable parts: an Attribute Key and an Attribute Value. For both keys and values, the following principle holds: they can use a self-contained terminology (i.e., Self-Explanatory for keys or Intrinsic for values), meaning that they only provide the information their name conveys, as opposed to linked terminology (i.e., Explained for keys or Extrinsic for values, termed here as Qualified Attribute Items), so that the meaning they convey is enriched by the connection (*external dependence*) with another FAIR dataset (see **I3**). Qualified attribute items form parts of IDs, which are special types of attributes issued by an Identification Service through an ID Registration relator (see **F1**).

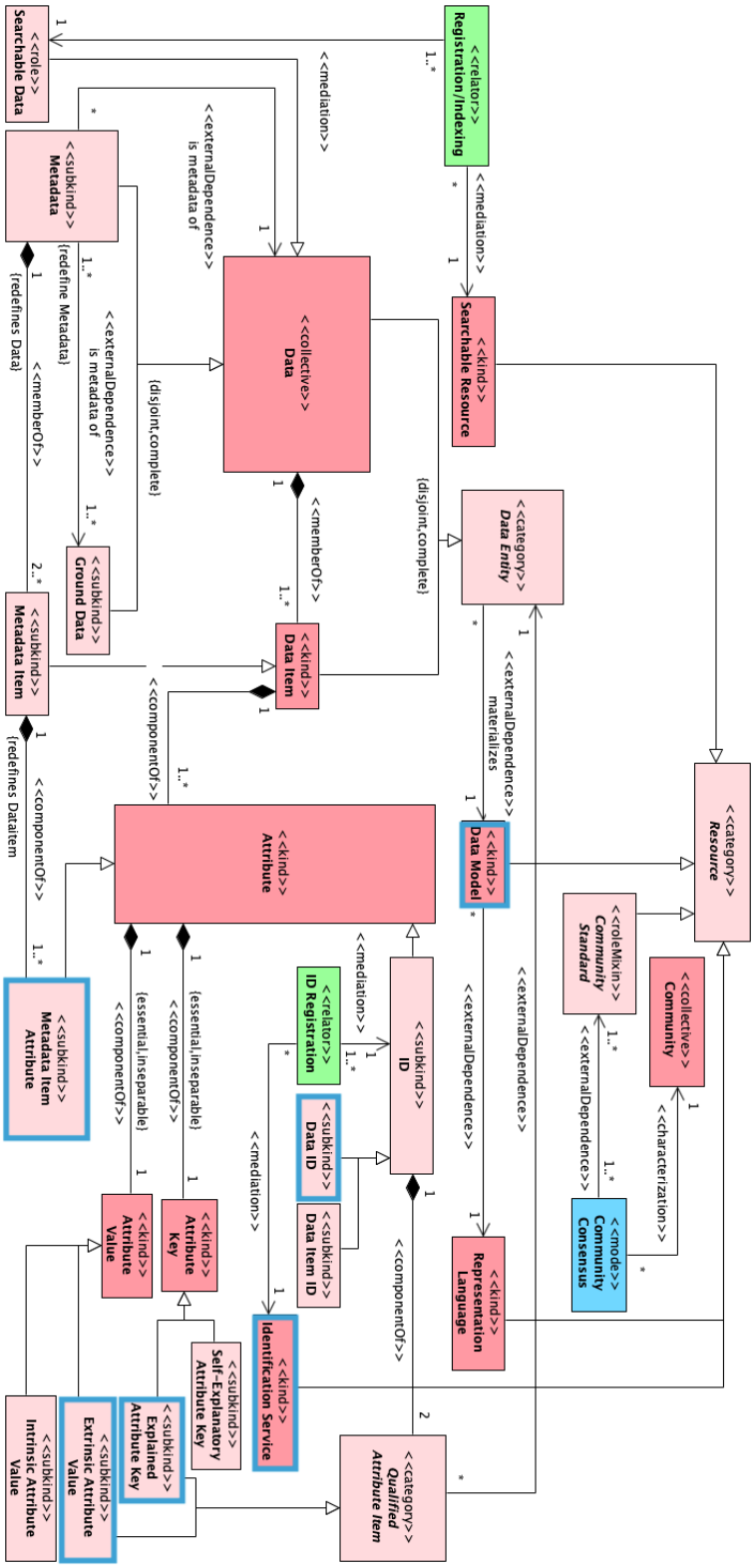

Fig. 1. OntoUML FAIR Principles Schema: module that represents Findability and Interoperability aspects.

Identifiers can then be used to identify (meta)data via these qualified references, thereby allowing data sources (through their metadata) to interoperate with (meta)data of other sources. This mechanism is made possible by means of explained keys and extrinsic values. The model contains a path that illustrates this recursive mechanism, which is left implicit in the FAIR principles (see **I3**). Note that, based on this representation, we do not need to explicitly model the 'controlled vocabularies' mentioned in **I2**, because they can be intended as represented by Data sources themselves, referenced by metadata qualified attributes. IDs are specialized into Data ID - identifiers of a dataset, as well as Data Item IDs, i.e., not only the whole collective can have identifiers, but single items can as well. Furthermore, as mandated by **F3**, a data ID must be represented as a metadata item attribute. Searchable Resource, Data Model, Representation Language, and Identification Service are all Resources and, hence, consensus and community standards can be established for all these types of entities.

Accessibility. The excerpt of the schema related to this principle is shown in Fig. 2. The retrieval of metadata should always be possible, according to FAIR (**A2**). Depending on the choices and policies of the data-creators, data can be (contingently) Accessible Data, which in turn can either be Open Data or Data with Restricted Access. Data is accessible if it has an identification scheme (see ID) and explicitly prescribed Data Accessibility Requirements defined as a contract (in the sense of [10]), but necessarily in a machine-readable format, i.e., as a machine-readable bundle of social and legal rights, obligations, powers, etc., connecting a given a data set with a Community. When data has restricted access, we use the Restricted Data Accessibility Requirements. Data accessibility requirements demand Data Access Protocol (see **A1.1**), which can be Authorization Protocols or Authentication Protocols, specifically included in restricted data accessibility requirements (see principle **A1.2**). Protocols (and specializations thereof) are Resources and, once more, they can be subject to a Community Consensus and then accepted as Community Standards (see **A1**).

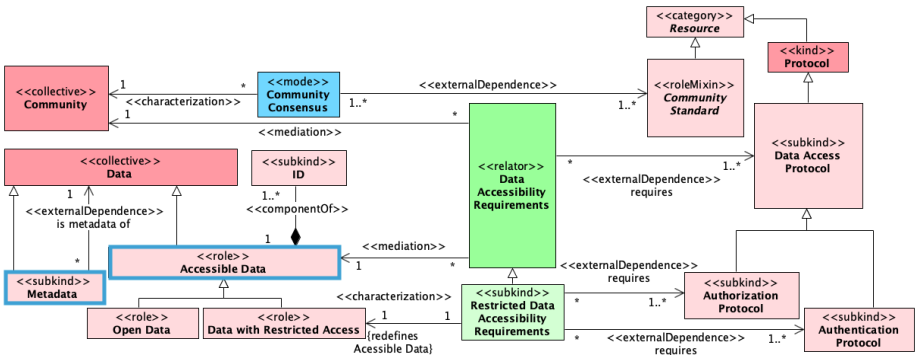

Fig. 2. OntoUML FAIR Principles Schema: module that represents Accessibility.

Reusability. The excerpt of the schema related to this principle is reported in Fig. 3. Here, a Rich Metadata is a collective of Rich Metadata Item Attributes (**R1**). The latter is a role played by Metadata Item Attributes when described following Community Standards for accuracy and relevance (see **R1** and **R1.3**). If a Metadata Item Attribute describes provenance information then it is also a Provenance Metadata Item Attribute (see **R1.2**), and metadata containing Provenance Metadata Item Attributes are considered to be Provenance Metadata. If a Provenance Metadata Item Attribute is richly described according to Community Standards then it is a Rich Provenance Metadata Item Attribute If all the constituents of a Provenance Metadata collective are Rich Provenance Metadata Item Attributes, then we have a Rich provenance metadata collective. One could argue that the above description does not fully characterize what is intended by 'rich' in this context. However, it does emphasize the role of community standards in this process. That is, the principles prescribed that - if Community Standards exist - they should be preferred to any other initiative with the same objective as the standard. Frequently, one standard does not cover all the needs for (meta)data, requiring the combination with other standards and/or approaches. (Meta)Data is considered here to be Reusable Data if it is described by Rich Metadata (including Rich Provenance Metadata, see **R1.2** and **R1.3**) and if we have explicit Data Usage Licences associated to it (see **R1.1**). The latter is a contract (again, ideally in the sense of [10]) directed towards a target Community.

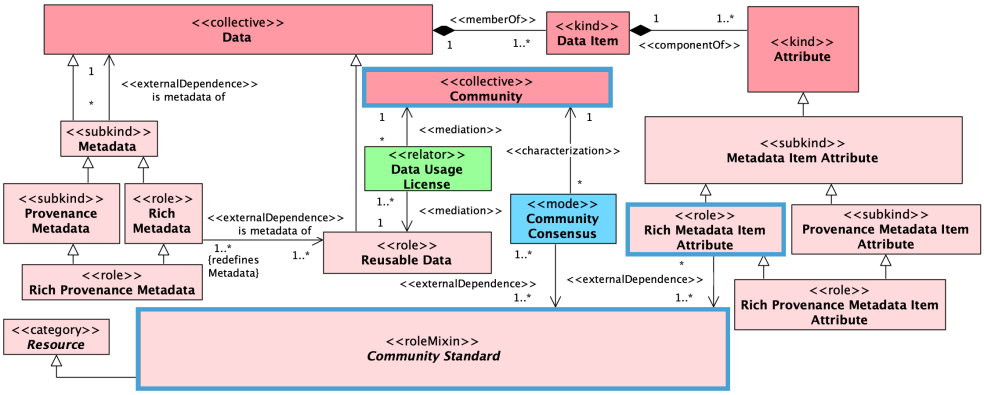

Fig. 3. OntoUML FAIR Principles Schema: module that represents Reusability.

The OntoUML FAIR Principles Schema as an ontology. The described artifact is a schema that results from a process of ontological analysis. We do not call it a FAIR ontology, to avoid terminological confusion. In the conceptual

modeling literature, the term ontology is used in three different ways, which are briefly discussed in the sequel.

In the first sense, an 'ontology' is a theory about what is assumed to exist by a representation artifact [15]. The process of ontological analysis here reveals the ontology behind the original description of the FAIR principles. The result of this analysis is explicitly captured in an OntoUML model, whose primitives make explicit the ontological categories of the involved domain notions, as well as the so-called *truthmakers* of the propositions constituting that description [12]. This is a process of explanation (of the ontology underlying a certain description) that is called *ontological unpacking*. The process of unpacking that we have employed here is described in [16,9,3]. In this sense, our proposed schema is a representation of the data ontology behind the FAIR principles.

In the second sense, an 'ontology' is supposed to be a "formal, explicit specification of a shared conceptualization" [4]. Given that (1) the FAIR principles themselves represent a shared conceptualization of data management guidelines; (2) they are explicitly represented in the OntoUML model/specification; and (3) OntoUML has a formal semantics [14], then the proposed OntoUML schema can be termed an 'ontology' (in the second sense).

Finally, there is a third sense in which an ontology is taken to be "equivalent to a Description Logic knowledge base" [18]. This is typically represented in the Web Ontology Language (OWL). One of us has argued in depth elsewhere why this is a problematic interpretation of the term [15]. In any case, the OntoUML tool set includes a fully automated approach for generating OWL specifications from OntoUML models [13]. So, an 'ontology' in this third sense can be automatically generated from the proposed schema.

4 Example Implementations

To illustrate the usage of our proposed ontological schema, from the expected benefits mentioned in Section 1, we discuss an example of data annotation with the schema (Section 4.1) and the use of our model for deriving prescriptive guidelines that operationalize the principles in a specific case (Section 4.2):

4.1 Semantic Annotation

Here we present an example concerning the genomic information related to the BRCA1 gene, which produces proteins that help repair damaged DNA. Certain harmful mutations in this gene increase the risks of several cancers, most notably breast, and ovarian cancer. Its information can be obtained from the Gene Database of the RefSeq data source [24]. From `https://www.ncbi.nlm.nih.gov/gene/672` we downloaded three files, whose partial information is represented in the Object Diagram in Fig. 4.

- the Ground Data instance `gene.fna`, i.e. a FASTA file storing DNA sequence stretches on different Data Items (rows), with two attributes: 1) a content-oriented Attribute, with a Self Explanatory Attribute Key

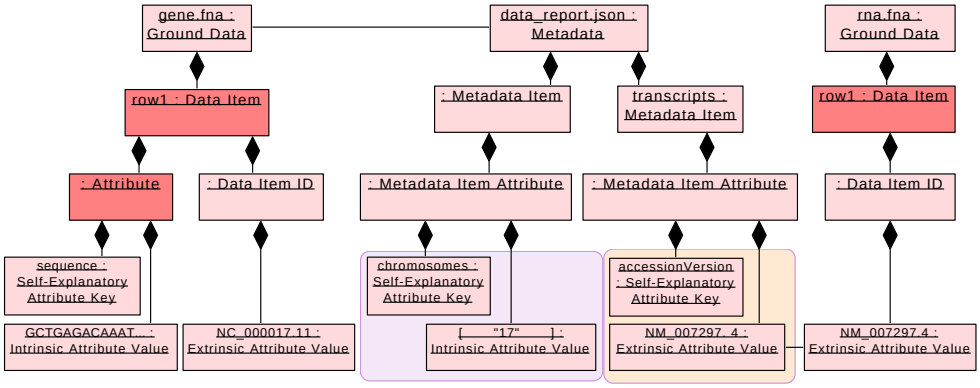

Fig. 4. Object Diagram of the OntoUML FAIR Principles Schema, for RefSeq data.

called 'sequence' and a very long Intrinsic Attribute Value starting with 'GCTGAGAC...'; 2) an identifying attribute (the Data Item ID) following the FAIR principle **F3**, holding an Intrinsic Attribute Value 'NC_000017.11'.

- the Ground Data instance `rna.fna`, containing RNA sequence stretches, with a Data Item identified by the Data Item ID 'NM_007297.4'.
- the Metadata instance `data_report.json`, with (1) atomic Metadata Items with one JSON element (Metadata Attribute), containing information, e.g., on the human organism to which the gene belongs (not shown in the diagram) or on the chromosome (key) and its number '17' (value). See Fig. 4 (purple background); (2) composite/nested Metadata Items, e.g., about the transcripts of the gene. Each transcript item includes several Metadata Item Attributes, e.g., name, length, ensemblTranscript (not shown), and accessionVersion. The accessionVersion is a Self-Explanatory Attribute Key corresponding to the 'NM_007297.4' Extrinsic Attribute Value (see Fig. 4, orange background). This is a Qualified Attribute Item forming the ID of the `rna.fna` Ground Data mentioned above. It is important to note that the extrinsic 'NM_007297.4' value allows us to connect the data attributes of `rna.fna` with the metadata attributes of `gene.fna`.

4.2 Refining and Operationalizing Guidelines

The FAIR principles were originally defined in an abstract manner thus hampering its direct operationalization. The ontological analysis and resulting schema conducted here can be systematically employed to refine and propose concrete guidelines for the operationalization of principles. Due to space limitations, we will illustrate this process here by addressing only certain aspects of a few principles.

First, consider **R1.3**. The model of Fig. 3 makes it explicit that community standards are established by the collective belief of a community (a community consensus). This consensus can be established with respect to resources in general, not only with respect to the rich description of metadata and metadata item attributes (see Figs. 1–2); for example, data models, representation languages, as well as data access protocols. In fact, this notion of employing resources that are deemed consensual in the collective belief of a community, also appears in a less obvious way when referring to protocols (see "standardized communication protocols" in **A1**), to representation languages ("broadly applicable languages" in **I1**), and to data models ("vocabularies that follow FAIR principles" in **I2**) - the original proposal explicitly refers to these as "community-endorsed vocabularies" [33], for example. Moreover, the community whose consensus one needs to follow is the target community towards which Data Accessibility Requirements and Data Usage Licenses are directed by Data entity creators. In fact, our analysis makes the centrality of communities and their collective beliefs clear, despite the fact that the term community is used only once in the description of the principles (**R1.3**).

As a result of this analysis, one can refine the aforementioned principles. It is important to make explicit that, when creating data entities, data creators must identify target communities to which accessibility requirements and usage licenses shall be directed, and which standards regarding protocols, data models, representation languages, and approaches for metadata description need to be properly identified and adopted.

5 Discussion

Several aspects of our work require discussion. First, selected aspects of the FAIR principles need additional effort to achieve a more precise specification. We highlighted the corresponding areas in Figs. 1- 3 with a light-blue outline. Second, the FAIR principles, as currently stated [33], provide asymmetric definitions. Third, there is a need for the identification of communities before community standards can be created and applied. Finally, the term 'rich' seems too broad to be interpreted properly within a scientific context.

1) Under-specified areas.
In Fig. 1, we highlight that the Identification Service shows another crucial part of the FAIR ecosystem, as mandated by principle **F1**. Schwanitz et al. [29] reviewed 80 representative databases employed in research on low carbon energy using the automatic evaluation framework proposed by proponents of FAIR principles [34]. However, none of these databases complied with **F1**, which means that data identifiers are not persistent in any of these cases.

The Data ID should be represented as a Metadata Item Attribute according to **F3**. While this is currently implemented by some genomic data sources (see [11,2,5]), it is less apparent in others [25,21].

The Data Model is a fundamental entity to comply with principle **I1**. Data models currently face many challenges, as highlighted in the Dutch initiative on

FAIR Genomes [31]. This includes the non-negligible problem of dealing with the updating strategy of data models, which must remain compliant with the domain to be represented as well as with the followed principles.

Moreover, the presence of Explained Attribute Keys and Extrinsic Attribute Values is tightly related to principles **I2** and **I3**. This highlights a very important requirement that is not yet broadly accepted in scientific data sources. For example, in The Cancer Genome Atlas (part of Genomic Data Commons [11]), clinical and biospecimen metadata are arranged according to a complex XML schema that contains self-explanatory keys of difficult interpretation, without following a controlled vocabulary. Examples include the: 'clinical.ablation.ablation_performed_indicator' or 'biospecimen.bio.bcr_analyte_barcode'. Instead, in Roadmap Epigenomics [21], tissues are described with values decided in-house by the data curation team of the source (see `https://www.roadmapepigenomics.org/data/tables`). Possible values include 'GI-COLON' or 'MUSCLE-Adult', which can lead to different interpretations when the corresponding genomic data is integrated with similar datasets, thereby hindering the possibility of their joint use. Such intrinsic attributes would clearly benefit from becoming extrinsic, using, for example, well-recognized ontologies for Anatomic sites (e.g., UBERON [22]).

According to the **A2** principle and, as observed in Fig. 2, when data is inaccessible (i.e., the complement class of Accessible Data in that model), Metadata should be accessible through the corresponding event. However, there are still several examples in the literature where this does not occur; e.g., in the energy domain [29] or the genomics domain [23].

Fig. 3, highlights Community Standards (discussed by **R1.3**). This is still a controversial concept because, in a given well-defined domain, it is not clear who should define the community standards for that domain. Large initiatives are starting to recognize these issues. For instance, Uniprot [26] claims to be FAIR for all other principles except R1.3, because the recognized authority is not known within the field of proteins [7]). Similarly, this aspect also remains unresolved in the field of low carbon energy data after a systematic assessment [29].

2) Asymmetric definitions. There is an asymmetry in the 4 letters inherent in FAIR. Findable (F) is fundamentally based on technology, whereas Interoperability (I) is very conceptual. In OntoUML we had difficulties modeling interoperability, highlighting that it is highly under-specified or under-formalized in the current FAIR principles.

3) Unclear community roles. In OntoUML, we modeled Community Standard as a role played by resources of different kinds when enriching data item attributes (turning them into Rich Metadata Item Attribute, i.e., a metadata attribute that is complete, generous, and explanatory enough to be adopted and reused by anyone in the Community). Our representation should facilitate a discussion on the characteristics of a good community standard: *Who should define it? How should it be formalized, communicated, and enforced?*

4) How rich is 'rich'? The 'rich' adjective could refer to a property of the meta-schema or something assigned with a relationship to the community that

defines or uses the metadata. If a rich metadata attribute is shared by multiple communities, is it more relevant than one that is only considered by one? A challenge is how to deal with discrepancies between communities from the same domain.

Aspects 3) and 4) are tightly related. We consider an 'unfair' situation to be one in which there are no community standards or it is not possible to clearly and unambiguously identify them. As a result, it might not be possible to define rich metadata attributes. Consequently, it becomes difficult to ensure that metadata has the required minimum quality for FAIR (specifically, for R1.3, but also for Findability and Interoperability-related principles). Overall, our results suggest the need for more efforts to define and agree upon community standards.

6 Conclusion

The adoption of FAIR principles for datasets is important and well-recognized by data scientists. In this research, we proposed an OntoUML FAIR Principles Schema to extend work on the adoption of FAIR principles by providing an ontologically grounded schema. The schema was applied, and its results shown to be effective, thereby establishing support for the implementation of FAIR principles. Further research is needed to apply this schema to other applications and refine it to support all aspects of the FAIR principles.

References

1. Ammar, A., et al.: A semi-automated workflow for FAIR maturity indicators in the life sciences. Nanomaterials **10**(10), 2068 (2020)
2. Bernasconi, A., et al.: META-BASE: a novel architecture for large-scale genomic metadata integration. IEEE/ACM Transactions on Computational Biology and Bioinformatics **19**(1), 543–557 (2022)
3. Bernasconi, A., et al.: Semantic interoperability: ontological unpacking of a viral conceptual model. BMC Bioinformatics **23**(11), 491 (2022)
4. Borst, P., et al.: Engineering ontologies. International journal of human-computer studies **46**(2-3), 365–406 (1997)
5. Buniello, A., et al.: The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Research **47**(D1), D1005–D1012 (2018)
6. Devaraju, A., et al.: An automated solution for measuring the progress toward FAIR research data. Patterns **2**(11), 100370 (2021)
7. Garcia, L., et al.: FAIR adoption, assessment and challenges at UniProt. Scientific data **6**, 175 (2019)
8. García S, A., et al.: Empirical Assessment of an Ontological Model of the Human Genome. In: Advances in Conceptual Modeling. pp. 55–65. Springer (2022)
9. García S. A. et al.: An Ontological Characterization of a Conceptual Model of the Human Genome. In: CAISE'22 Forum. pp. 27–35. Springer (2022)
10. Griffo, C., et al.: Conceptual modeling of legal relations. In: International Conference on Conceptual Modeling. pp. 169–183. Springer (2018)
11. Grossman, R.L., et al.: Toward a shared vision for cancer genomic data. New England Journal of Medicine **375**(12), 1109–1112 (2016)

12. Guarino, N., et al.: "We need to discuss the Relationship": Revisiting Relationships as Modeling Constructs. In: Proc.CAISE'15. pp. 279–294. Springer (2015)
13. Guerson, J., et al.: Ontouml lightweight editor: a model-based environment to build, evaluate and implement reference ontologies. In: IEEE EDOCW'15 (2015)
14. Guizzardi, G.: Ontological foundations for structural conceptual models. CTIT, Centre for Telematics and Information Technology (2005)
15. Guizzardi, G.: Ontology, ontologies and the "i" of fair. Data Intelligence **2** (2020)
16. Guizzardi, G., et al.: Ontological unpacking as explanation: the case of the viral conceptual model. In: Proc. ER'21. pp. 356–366. Springer (2021)
17. Guizzardi, G., et al.: Ufo: Unified foundational ontology. Applied Ontology **17**(1), 167–210 (2022)
18. Horrocks, I., et al.: From shiq and rdf to owl: The making of a web ontology language. Journal of web semantics **1**(1), 7–26 (2003)
19. Jacobsen, A., et al.: FAIR principles: interpretations and implementation considerations. Data intelligence **2**(1-2), 10–29 (2020)
20. Kersloot, M.G., et al.: Perceptions and behavior of clinical researchers and research support staff regarding data FAIRification. Scientific Data **9**, 241 (2022)
21. Kundaje, A., et al.: Integrative analysis of 111 reference human epigenomes. Nature **518**(7539), 317–330 (2015)
22. Mungall, C.J., et al.: Uberon, an integrative multi-species anatomy ontology. Genome biology **13**, R5 (2012)
23. Nayar, P.G., et al.: CardioGenBase: a literature based multi-omics database for major cardiovascular diseases. PloS one **10**(12), e0143188 (2015)
24. O'Leary, N.A., et al.: Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Research **44**(D1), D733–D745 (2016)
25. 1000 Genomes Project Consortium: A global reference for human genetic variation. Nature **526**(7571), 68 (2015)
26. The UniProt Consortium: UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research **49**(D1), D480–D489 (2021)
27. Ruy, F.B., et al.: From reference ontologies to ontology patterns and back. Data & Knowledge Engineering **109**, 41–69 (2017)
28. Sansone, S.A., et al.: FAIRsharing as a community approach to standards, repositories and policies. Nature biotechnology **37**(4), 358–367 (2019)
29. Schwanitz, V.J., et al.: Current state and call for action to accomplish findability, accessibility, interoperability, and reusability of low carbon energy data. Scientific reports **12**, 5208 (2022)
30. Bonino da Silva Santos, L.O., et al.: FAIR Data Point: A FAIR-oriented approach for metadata publication. Data Intelligence (2022)
31. van der Velde, K.J., et al.: FAIR Genomes metadata schema promoting Next Generation Sequencing data reuse in Dutch healthcare and research. Scientific data **9**, 169 (2022)
32. Verdonck, M., et al.: Comparing traditional conceptual modeling with ontology-driven conceptual modeling: An empirical study. Information Systems **81**, 92–103 (2019)
33. Wilkinson, M.D., et al.: The FAIR Guiding Principles for scientific data management and stewardship. Scientific data **3**, 160018 (2016)
34. Wilkinson, M.D., et al.: Evaluating FAIR maturity through a scalable, automated, community-governed framework. Scientific data **6**, 174 (2019)