Beyond Simple FAIR Principles for Ontologies and Semantic Resources

Grounding Rich, Meaningful Metadata

Gary Berg-Cross[1], Sage Arbor[2]
[1]ESIP Semantic Harmonization, [2]Duke Clinical Research Institute

Abstract

Digital forms of metadata such as controlled vocabularies, taxonomies and conceptual models play an important role in ensuring that data satisfy the principles of findability, accessibility, interoperability, and reusability (FAIR). In turn, metadata also relies on semantic artifacts called formal ontologies to make metadata computer processable. Ontologies make metadata semantically rich with axiomatized definitions that represent useful meanings, but adapting simple ideas of FAIRness for a broad class of digital objects called semantic resources, especially ontologies, raises a number of semantic issues. The focus here is on issues involving community standards for rich metadata and adequate grounding of these with meaningful semantics. This comes despite the fact that in many ways, ontology developments have preceded, and proceeded well beyond, simple FAIR principles. We illustrate the value of community standards by the development of capabilities to document ontology modules sharing a common framework. As part of grounding semantics, we suggest a useful direction is to capture the form of axiom patterns using common ontology design patterns, which are themselves grounded in foundational concepts.

1. Introduction

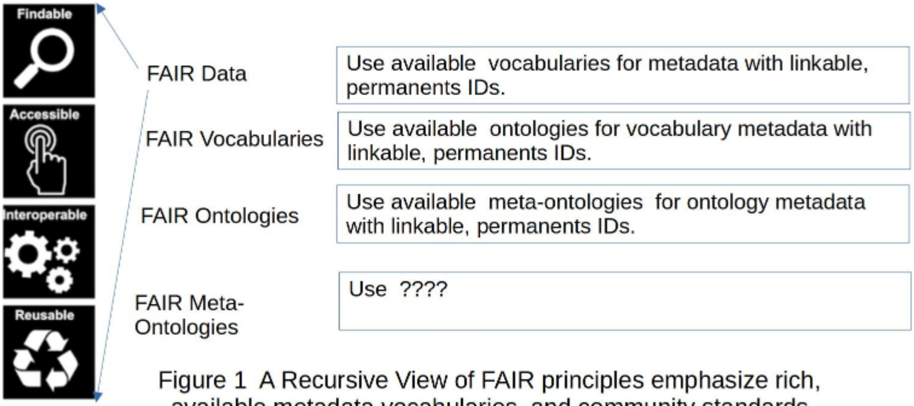
SCIENCE, ESPECIALLY OPEN SCIENCE IN A DIGITAL AGE, is increasingly data driven, relying on integrated data access via the Web. This open science strategy to use information artifacts has long been challenged by the heterogeneity of big digital data and has led to increased reliance on metadata to support finding, understanding and using relevant managed digital data. As an aid to data management, the FAIR initiative (Wilkinson et al., 2016), with four foundational principles for Findablity, Accessiblity, Interoperablity, and Reusablity, has gained significant traction. FAIR principles provide, for example, some guidance on metadata practices to support open data access and operation (Blomberg et al., 2016). An important feature of the FAIR principles is to go beyond simple metadata using key-value pairs that describe basic characteristics found in typical, published datasets. Examples of simple metadata are tags for titles, identifiers, text descriptions, keywords, publisher name, date of publication, or terms of use. Databases such as the Disease Database (2022) are illustrative of using this simple type of metadata to document data.

The FAIR foundation and more detailed guiding principles grew in part out of the recognition that the vocabularies needed for metadata should be standardized and further that some semantics is needed to enable machine-actionability (i.e., the capacity of computational systems rather than humans to find, access, interoperate, and reuse data). Partly in response to the useful role of FAIR principles many diverse semantic resources and knowledge graphs (KGs) have been developed across the domains. These domains are as varied as the biomedical and earth sciences. KGs are very popular as useful integration artifacts because they in effect turn related pieces of data into actionable knowledge units (De Smedt et al., 2020).

To help with such things as data integration used in KGs, FAIR includes rich metadata as a central concept as well as some metrics to help objectively score FAIRness. This highlights the dependency of FAIR data on appropriately rich, community agreed upon metadata. Such metadata starts with FAIR vocabularies that are linked and have formal semantics to allow automated processing (Figure 1). Another central feature, illustrated in Figure 1, is that FAIRness is recursive with the role of FAIR metadata at its level requiring some more fundamental grounding in a richer level.

So FAIR vocabularies need to be grounded in FAIR ontologies. These are the ones obeying FAIR principles like findability. But how is FAIRness assessed?

As an aid FAIR provides some best practices of metrics for data/informational resources (Crosswell, 2022). Metric tools, like self-assessment instruments and validators, were developed early on (Wilkinson et al., 2018; Devaraju et al., 2021) to generate a global FAIRness score. But as work on FAIR has progressed it was recognized that the original technology neutral concept of self surveyed "metrics," involved a degree of subjective interpretation. This threatens the idea of adequate grounding of metadata for both machine processing and human understanding. The response has been to develop several more objective measures called "Maturity Indicators" (FAIR Maturity Indicator, 2019). Essentially maturity indicators



Use available vocabularies for metadata with linkable, permanents IDs.

Use available ontologies for vocabulary metadata with linkable, permanents IDs.

Use available meta-ontologies for ontology metadata with linkable, permanents IDs.

Use ????

Figure 1 A Recursive View of FAIR principles emphasize rich, available metadata vocabularies and community standards

describe the various components of FAIRness in operational ways that can be objectively evaluated by semantic technologies tools (RDA, 2020). For example, the first maturity indicator for findability, checks that a repository implements accessible, structured metadata to enhance its discoverability on the Web. This is essentially an existential check that a globally unique and permanent identifier (GUID) has been used to name a data element and that an associated link can be resolved. The presence of any Linked Data that can be found for a data element is then considered an objective success (FAIR Maturity Indicator, 2019). A similar idea is adopted for metadata. The metadata maturity indicator is an existential check that metadata follows some “community-defined model.”

In both cases there is a degree of simple objectivity, but, on reflection, one might ask whether we have measured everything of interest and in particular the meaning understood for vocabulary terms serving a metadata role. There is, for example, no unique definition for “soil” or ”soil density,” and so there may be many different, but unique, GUIDs for a named term in a vocabulary, defined in a vocabulary glossary and/or data using the vocabulary label “soil.” Similarly, a community standard may exist for defining particular metadata items, but the quality of this standard, especially older standards that are not actively maintained, may vary in important ways.

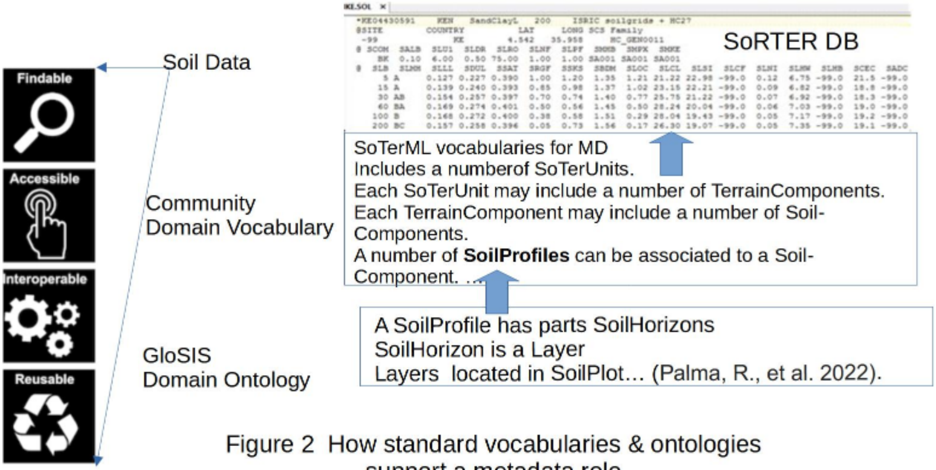
As an example, Maron and Feinberg (2018) discuss the importance of the way that standards creators, like the Dublin Core standard, understand what it means to “adopt a standard.” This is different from the way that web publishing platform implementations share digital collections as well as how regular users understand what it means to “adopt a standard.”[1] A key point here is that there can still be an uncomfortable degree of subjective understanding of maturity indicators and what a particular guideline means. This highlights the centrality of semantic meaning in grounding FAIRness. Unfortunately at the domain level (e.g., what is soil density?), the required data vocabularies to use for some domain’s metadata are currently little more than lists, while others contain arbitrary definitions. This lack of domain definitions that have good quality and are accepted by the community is clearly a problem for FAIRness. Further, vocabularies limited to human readable text do not follow the encouraged FAIR practice for metadata schemes, implied in Figure 1, which is to use a formal, accessible, shared, and broadly applicable language for knowledge representation based on semantic models such as formal ontologies (Guizzardi, 2020). If, for example, vocabulary can be described with measurable units (e.g., soil density as kg/m^3), more robust universal definitions are possible. However many vocabulary words do not encompass such units in their meaning.

Ontologies, with standardized lexicons that include information about how items are categorized and related to one another, play a role both in building KGs and in supporting FAIR data principles. Such ontologies can provide semantically defined vocabularies usable for what FAIR calls rich metadata annotations for data. The relations in Figure 2 illustrate how vocabulary standards, such as SoTerML (Pourabdollah et al., 2012) and ontologies, such as GloSIS (Palma et al., 2020), further enrich the data-metadata associations by embedding a particular concept within a related conceptual pattern or schema.

It has long been understood that some degree of useful semantic interoperability can be achieved by leveraging formal ontologies that capture meaningful conceptualizations and that can be shared as digital artifacts. As part of the Semantic Web initiative, Heflin and James Hendler (2000), for example, described the challenge of meaning-based data integration:

---

[1] A sign of how difficult it is to maintain links is the fact that FAIRsharing’s list (https://fairsharing.org/standards/identifierschema) of community-recognized identifier schemas has a broken link.



SoTerML vocabularies for MD
Includes a numberof SoTerUnits.
Each SoTerUnit may include a number of TerrainComponents.
Each TerrainComponent may include a number of Soil-Components.
A number of **SoilProfiles** can be associated to a Soil-Component. ...

A SoilProfile has parts SoilHorizons
SoilHorizon is a Layer
Layers located in SoilPlot... (Palma, R., et al. 2022).

Figure 2 How standard vocabularies & ontologies support a metadata role

> To achieve semantic interoperability, systems must be able to exchange data in such a way that the precise meaning of the data is readily accessible and the data itself can be translated by any system into a form that it understands.

However, the Semantic Web's vision of easy interoperability has not been widely achieved in the intervening years. While simple conceptual definitions such as seen in Figure 2 can serve as links between different data, there is no one, master, ontologically formalized conceptualization. Rather there are as many ontologies as there are vocabularies defining conceptualizations of data in different ways. This reflects that fact that humans understand and define a concept, such as with the label "soil," in many different ways. Thus ontologies that are structured using vocabulary definitions also come to reflect different understandings. While the word "soil," for example, is conversationally used and understood in what seems a common, conceptual space there is no universally recognized, precise definition. Contexts like agriculture evoke different conceptualizations of soil because they emphasize things like manure as a constituent beyond the more typical constituents of sand and clay. Vocabularies, like ontologies, may have implicit scope, but still may be understood by different people based on their role relations to the domain being formalized. Even more conceptual differences can be found, including the implicit meanings evoked by the topic of "soil liquefaction" or "soil contamination" (Todd-Brown and Berg-Cross, 2022).

Recognizing that meaning can change subtly in the relations found in different contexts encourages the use of supplementary semantic methods which go beyond creation of a simple conceptual model. These include standardizing and aligning definitions to help clarify extant conceptualizations, which can in turn support finding and using appropriate labels backed by community standard definitions for data. Thus connecting a basic concept of "soil" with usual constituents to a sub-concept of "toxic soil" with hazardous ingredients allows greater coverage across the wide conceptual space that people understand about soils.

So a simple application of ontologies does not solve all the problems of data and meaning with respect to a domain. Thus researchers have considered the FAIRness issues for ontologies themselves. This makes sense since, as seen in Figures 1 and 2, FAIRness applies to metadata and some of that FAIRness relies on ontologies. However, some problems for ontology FAIRness are illustrated by a recent survey on disaster management ontologies by Mazimwe et al (2021). They performed FAIRness tests using published information available as part of a review and found that FAIR principles were seldom followed for many of the ontologies analyzed. Surprisingly the average Findability[2] using the four FAIR Finding criteria was only 1.8%, and the average Accessibility, which is provided by functioning APIs, was only 5.8%. Only 4.3% of the retrieved ontologies provided details about explicit mapping/correspondences between ontologies. Metadata schema issues were also reported by Mazimwe et al (2021) such as a failure to use standard vocabularies to describe semantic artifacts. And rather surprisingly 90.9% of URLs provided for the artifacts did not conform to the well understood principle of uniqueness and persistence of links.[3] Some of the lack of resolvable URLs may reflect "link rot" and the fact that older (pre-2015 and hence pre-FAIR principles) ontologies were included. Thus for a number of reasons it seems useful to examine some issues about FAIR ontologies, especially semantic and community issues that may involve more than simple ideas of FAIRness.

---

[2] Findability involves assigning unique and persistent identifiers to a digital object, and describing them with rich metadata that enable their indexing and discovery.

[3] The first access principle requires that (meta)data are retrievable by their identifier using a standardized communications protocol.

This paper discusses examples of community standardization efforts, technologies and methodologies for developing FAIR ontologies and takes a deeper look at ontology metadata and how harmonization across the semantic spectrum is needed to facilitate semantic interoperability. The remainder of the paper is structured as follows. Section 2 provides context and illustrates some work and issues arising from a simple application of FAIR principles to ontologies and the automation of meaningful metrics. In Section 3 we provide some examples of independently developed guidelines for metadata documentation of ontology characteristics and content. Section 4 presents a biomedical example of FAIR ontologies in the complex COVID-19 domain and illustrates the value of a coordinated, community effort to build related and interoperable ontologies. Section 5 concludes the paper by examining the problem of grounding metadata so that it has the rich semantics that is necessary for achieving the FAIR principles and beyond.

2. Applying FAIR Guiding Principles to Ontologies

As context, some examples of FAIR ontology work is provided to help understand some of the issues involved in finding, accessing and sharing ontologies. We start with the idea of rich metadata used for ontologies. We leverage rich metadata guidelines and methods that were developed for the FAIR principles before looking at semantics beyond FAIR.

Ontological challenges relating to FAIRness, for whatever purpose, start with finding one or more relevant ontologies and drilling down to concept(s) within the ontology[4]. Providing rich metadata describing ontologies in a meaningful way is essential for finding an ontology[5]. We can see how this works from a FAIR perspective by simply replacing the concept of “data” with “ontology” in FAIR guidelines (Poveda-Villalón, María et al., 2020). This give us an aspirational start on finding relevant ontologies and their conceptual elements using guiding principles:

FO1 Ontologies (and their elements) are assigned a globally unique and persistent identifier.

---

[4] Finding and accessing knowledge are closely related in deciding which ontologies to use and are discussed together in this subsection.

[5] There is an obvious degree of recursion here since metadata describing an ontology must itself be FAIR and such vocabularies should be grounded in fundamental concepts about ontologies.

FO2 Ontologies (and their elements) are described with “rich” metadata (meta-ontologies[6]).

FO3 Ontologies (and their element’s) metadata clearly and explicitly include the identifier for the ontology/elements they describe.

FO4 Ontologies (and their elements) are registered or indexed in a searchable resource.

FO1 reflects principles like universal identifiers and resolvable linkages. Ontologies have long included some such practices to help find and access their contents. Guidance to find elements was part of the earlier Linked Data principles growing out of the Semantic Web in 2006. Before the widespread acceptance of FAIR principles for data research there were Linked Data guidelines along with experience publishing ontologies on the Web using permanent identifiers and making them available through the HTTP protocol (Janowicz et al., 2014). Despite this, as Mazimwe et al (2021) have shown, even easy practices such as resolvable GUIDs may not always be followed.

FO2 requires metadata about an ontology, most importantly what its content is. Part of the early work with ontology repositories demonstrated the lack of community agreed upon metadata standards for fully describing even the basic content of ontologies. Dutta, Nandini and Shahi (2015) observed, for example, that while the majority of ontology libraries/repositories used the term “author” to capture authoring information of an ontology, some of the early libraries used the term “creator.” This seems simple to fix with linguistic concepts like “synonyms.” But other meta-aspects of rich metadata, such as how to describe an ontology’s content and structure to find a relevant ontology, are more difficult to agree on and implement.

Richness can be subjective and thus some of what humans find satisfyingly rich based on available background knowledge can be hard to formalize for computer processing. Some knowledge represented in, say an ontology schema, might provide a useful index to find a relevant ontology. But there is yet no universally agreed upon schema to represent the complexity of a particular ontology’s content. Instead, any checking to find

---

[6] There is an obvious recursion in richly describing ontologies by other ontologies having a meta-ontology role.

an appropriate ontology (or another semantic resource like a controlled vocabulary) typically requires considering characteristics, such as quality of defined meanings. These are characteristics that make an ontology relevant, trustworthy and useful.

Simple ontology search tools do exist to use for some domains such as biomedical topics (The Ontology Lookup Service, 2022). But they are built to find a relevant concept based on a label like "disease" or "soil." And while a list of candidates may be provided by such a simple search, further analysis of candidate ontologies is usually needed. For example studying taxonomic structure is important, which in turn requires access and analysis (Jacobsen et al., 2020).

Use of metadata for interoperability is even more challenging than finding separate but related ontologies. To start, we require APIs between the repositories containing the ontologies. Moreover, these APIs may need to access other ontological knowledge, possibly in other repositories, because ontologies, like data and metadata, are distributed. How does metadata tell us about what axiomatized information can be shared? How does an automated operation tell that two definitions are equivalent or even compatible?

We can provide a start on simple guidance for this type of operation by simply translating three FAIR data interoperability principles to ontologies and see where they lead. We then have:

IO1 Ontology metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation across different ontologies. As part of a Semantic Web vision, the original interoperability guidance was simple. The idea was to use metadata vocabularies that conform to the major available logic-based knowledge representation languages (Berners-Lee et al., 2001). Unfortunately this ignores the content of what is represented.

IO2 Ontology metadata use controlled vocabularies that follow FAIR principles.

IO3 Ontology metadata includes qualified references to other ontology metadata.

More experience with data interoperability has led to growing requirements beyond the guidance of IO1 to just semantic language standards or even a direct use of ontologies represented in these languages. However, the principle of "garbage in-garbage out" still applies, even if the representation language of an ontology allows automated processing. The quality of what is represented in an ontology is what is important here and not just the representational language.

Because there is no one master ontology to represent everything, there is a need to interoperate between ontologies, guided by what we know of them from IO2 and IO3. While there may be existential items that can be checked from a meaning point of view, this is shallow and not rich metadata. One can argue that we need more meta semantics as we move to consider how to support interoperability of related ontologies, which are increasingly stored, like data, in repositories. Deeper, rich documentation here involves relating concepts. These may be captured in ontology mappings, alignment and even harmonization. But this can be hard and a common approach is to simplify interoperability often to something like a word-based shallow alignment between ontologies. One example of simple alignment uses the Simple Knowledge Organization System (SKOS), which is based on labels and/or a simple thesaurus scheme. SKOS represents similarity of terms using linguistic or hierarchical representations like "broader term" or loosely associative ones like "related terms" as opposed to richer representations of semantic relations like "same class as." While the SKOS vocabulary is shallow, it is useful because it is formalized in a Resource Description Framework (RDF) processable format. This provides at least an entry step to an incremental semantic approach, which allows for subsequent work to expand the scope and to provide richer domain semantics (Berg-Cross, 2021).

3. Beyond FAIR to Richer Metadata Vocabularies for Ontologies

Ontological methods have something to contribute beyond the FAIR guidelines. Interestingly, standardizing metadata for ontologies has followed an incremental path to maturity that is a bit independent of, and partly precedes, the FAIR initiative. To some extent this independent path may have made ontology developers less motivated to use simple FAIR ideas for ontology development and documentation. Here we can agree with Amdouni, Bouazzouni, and Jonquet (2022) that we have yet to see a clear methodology implemented and tooled to automatically assess the level of FAIRness to ontologies. But in some ways, even without common, tooled methods, we have worthwhile resources to take a few steps toward rich metadata for ontologies.

An initial step towards a standardized metadata vocabulary was taken with development of the Ontology Metadata Vocabulary (OMV) by Hartmann et al. (2005). The work provided a broad, if not deep, view of key metadata requirements, which, in part, both prefigure FAIR principles by some 10 years (the ones overlapping with FAIR are emboldened below) and include semantic aspects beyond initial FAIR principles.

Eight factors were developed as listed below, and a few highlights are worth noting. One is that OMV includes factors like usability for humans. But as a formal semantic effort it also provide guidance on making the models processable. Another is that it includes an important distinction between the rationalized conceptualization underlying ontological development and the pragmatics of how an ontological model might be realized in ways that are helpful for a range of users. Another important idea is that it lists some minimal documentation of key information.

1. **Accessibility**. Ontology metadata must be accessible and processable for machines as well as humans.

2. Usability. The metadata should be understandable so that a majority of domain users should be able to readily apply metadata.

3. **Reuse**. Because ontologies are a (or the) core technology for the Semantic Web, its metadata should reflect key issues of the Semantic Web such as reuse and sharing of knowledge.

4. Conceptualization vs. Realization. Metadata must reflect a distinction between the underlying semantic conceptualization and its particular realization as a concrete ontology digital artifact.

5. **Interoperability**. Metadata interoperability is key and was defined as conforming to the major representation languages being used for Semantic Web applications in 2005, such as the Web Ontology Language (OWL).

6. Documentation. Metadata documentation should provide at minimum information about technical, statistical, accessibility, management information, etc.

7. Extensibility. Reflecting special user needs and the reality that on-

tologies are built incrementally as knowledge requirements and issues unfold, it is required that beyond such standard metadata, facts can be added and extended easily.

8. Expressiveness. Metadata must be expressive enough to represent all desired aspects, as presented above.

With the growing impact of ontology repositories between 2005 and 2014 the wide, principled scope of metadata, the proposed OMV needed more detail to guide behavior. OMV was followed by a more detailed de facto guideline called "Metadata for Ontology Description and publication" (MOD) (Dutta, Nandini and Shahi, 2015). This reflected the actual experiences to find and share relevant ontologies in repositories such as BioPortal. MOD's guidance concerns how to document ontology characteristics and contents so that they are easily identifiable and reusable for various knowledge engineering tasks. These documentation requirements were derived from a bottom-up survey of repository searches, which identified detailed and specific properties used in ontology repositories. As a result of their survey and analysis, Dutta, Nandini, and Shahi (2015) also proposed eight requirements for ontology documentation. These partially overlapped with OMV's metadata requirements (such as Extensibility and Usability):

1. Brevity: The vocabulary should consist of a minimal set of elements maintaining balance between necessity and sufficiency.

2. Clarity: The metadata elements must be well defined, and clear descriptions should be provided.

3. Simplicity: The vocabulary should be easy to use.

4. Authority: The vocabulary design should be based on a sound methodology in the sense that the inclusion of terms in the vocabulary are justified.

5. Standardization: The element names should be standardized. To confirm the standardization, the individual elements should be mapped with the existing standard vocabularies.

6. Extensibility: The vocabulary should be extensible.

7. Usability: The vocabulary should support the reuse of the described resources. In other words, the vocabulary should allow the creators/developers to highlight the usage and the quality of the resources

in a well-defined manner.

8. Interoperability: The vocabulary should be interoperable. That is, it should conform to the major knowledge representation languages then in use for Semantic Web (Berners-Lee et al., 2001) applications.

As with FAIR principles there are issues with objectivity of the guidance. Characteristics like "brevity" and "clarity" are comprehensible to humans, but there is no easy guidance to make the brevity of metadata quantifiable. From a process point of view, what are well defined and clear descriptions? Further, principles like standardization of vocabularies do not address the quality of a vocabulary. More immediately helpful to documenting ontologies with metadata was MOD's use of seven metadata facets that include aspects like a basis for *Authority*, along with *Rights* and *Licenses* needed for ontological use. MOD facets are populated using 15 classes[7] that are formalized and expressed in the Web Ontology Language and thus processable[8]. The classes were derived by analyzing top-level facets such as *Authority*. For instance, the *Authority* facet identifies the person/organization that created the ontology and/or who exercises control over an ontology, or originated an ontology document. MOD semantics considers both *Person* and *Organization* as classes and these are grouped under a general class called *Agent*. The *Environment* facet includes classes like OntologyTool, OntologyLanguage, and OntologySyntax (Dutta, Nandini, and Shahi, 2015).

Perhaps the most interesting top-level facet is *domain coverage*. This often describes technical metrics about the ontology, such as the number of classes and properties, and the scope, which include requirements and a standard practice for developing ontologies called community competency questions (Aminu et al., 2020). Documenting competency questions takes metadata well beyond simple FAIR guidance. While a pattern of classes might be used to define a domain, often this is shallowly abbreviated with a label such as "protein" or "soil." But if reference is made to an ontology, a processable definition with axioms can be available and instances of a domain can be defined using RDF schemas. This is much more in the style of rich metadata, and we can see below an example of

---

[7] A class is a conceptual collection of things that share common attributes.

[8] Note that this was done several years before the FAIR standard and reflects earlier emphasis from the Semantic Web.

an axiomatized definition in the Environmental Ontology (ENVO) using standard relations also found in various other ontologies.

For example, Soil
    is a type of environmental material
    has quality some porous
    has part some sand
    has part some organic material
    has part some mineral material
    has part some (chemical entity and (has role some nutrient))
    part of some pedosphere
    has part some clay
    has part some silt

MOD has been used successfully as part of ontology development, but the range of possible metadata growing out of MOD's suggestions is large. Agro-Portal, for example, now recognizes 346 properties from existing metadata vocabularies that could be used to describe different aspects of ontologies and other semantic resources (Toulet, Dutta, and Jonque, 2018). This is rich in number but not always in depth. So choosing which metadata vocabulary items to use can be an issue.

A further tuning of documentation requirements provides some idea of which elements are essential as part of publication and/or in ontology repositories. The Minimum Information for Reporting an Ontology (MIRO) initiative, associated with the Open Biomedical Ontology (OBO) Foundry, provides guidance for what ontology developers should document about an ontology (Matentzoglu et al., 2018). If MOD was aimed at improving ontology repositories, MIRO added a goal of what should be documented in scientific reports. The MIRO guidelines specify the level of importance using a "must," "should" or "optional" label for each element as an aid to improving the quality and content consistency of the information descriptions. This further constrains the variability in descriptions of ontologies. At a minimum, for example, developers must describe the development methodology, provenance and context of information being reused. MIRO defines 34 information items, some of which, such as "ontology name", "ontology license" and "ontology URL," are content-independent, i.e., do not depend on what an ontology is about. They are thus not really rich metadata, although they will have GUIDs that are resolvable.

The MIRO work is significant, but as noted by Amdouni, Bouazzouni, and Clement Jonquet (2022), MIRO properties are not explicitly aligned with FAIR principles and assessment methods. In reality MIRO's 11 guideline items for ontology content go beyond FAIR and have been of importance for supporting ontology standardization and interoperability. In particular two guidelines go deeper than FAIR and were considered very important for users. These are information about ontology relationships and about axiom patterns. Both can be seen in the example of an axiomatized definition for soil. Guidance for relations and patterns reads as follows:

> **E.9** Ontology relationships and properties used in the ontology must be documented.
>
> **E.10** Axiom patterns must be documented. An axiom pattern is a regular design of axioms or a template for axioms used to represent a category of entities or common aspects of a variety of types of entities. The previous example of an axiom pattern used to define "soil" illustrates this.

In ontology tools like Protégé one can see and select from defined relations needed to implement E.9. Examples are *part*, *participation* and *composition* relations. In the OBO Foundry, for example, one may see a list of available relations stored in the Relations Ontology module (OBO Relation Ontology, n.d.). A *participation* relation, for example, has sub-types such as *input_of* to further specify the nature of participation. Metadata documentation of relations and axioms makes it possible to understand how concepts are conceptualized and represented. Documenting axiom patterns with a canonical set of Ontology Design Patterns takes one even deeper towards ontology interoperability, since the use of a common design based on good practices provides a common point for both development and inter-operation.

To help start ontology development one can search various catalogs of ontology design axiom patterns. These can be used as starting points for ontology development (ODP, 2022). Ontology design patterns are largely hand crafted because computationally extracting axiom patterns from an extant ontology remains difficult. However, early attempts have been made, such as extracting syntactic regularities from ontologies as proxies for axiom patterns (Mikroyannidie et al., 2011).

More recently, the Shapes Constraint Language (SHACL) standard and SHACL shape graphs provide a description of data graphs satisfying particular configuration conditions. This means one can define model shapes using vocabularies that are themselves defined and expressed in RDF graphs. In turn SHACL-defined shapes can be tied to the shape of ontological concepts and ontology design patterns because particular ontology design pattern axioms describing conceptual constraints on instances operate in a similar fashion to SHACL-generated constraints. Thus, closing the loop, an argument has been made that the structure of ontology design patterns might be used as an empirical basis for SHACL shape graphs (Pandit et al., 2018). This might afford some automated documentation of axiom patterns (Cimmino et al., 2020; Blomqvist et al., 2021) and provide a deeper basis for objective and processable axiom patterns used to describe ontology concepts.

An important distinction among ontologies is whether the ontology is intended to deal primarily with concepts or with data. The former kinds of ontologies will contain logical statements framed in terms of domain attributes and how they define the meanings of things in that domain (Bennett, 2021). These logical statements can be used by humans to understand things in the domain and to explain more complex collections of inter-related things. When things in a domain need to be represented with data, the ontology is often known as an “operational” or “application” ontology. Since the FAIR principles are concerned with data, it would seem that the only ontologies that would matter for FAIR are ones that deal with data. However, further analysis suggests that there may be different styles of ontology that deal with operational data. An ontology for integrating multiple sources of data may need to have more semantically nuanced distinctions to deal with the different ways those data sources reflect conceptualizations of the world. On the other hand, an ontology for reasoning over data (e.g., in a knowledge graph) would typically be simpler (Baclawski et al., 2021). Furthermore, even an ontology that is purely conceptual must still be represented using data (e.g., in OWL or first-order logic), and so the FAIR principles are relevant to the ontology itself.

There are other differences between operational and concept ontologies. For example, an operational ontology need not use a full foundational ontology to partition its world. It would also typically have fewer relationships, with little or no use of constraints such as property domains and ranges. By contrast, a concept ontology would have a richer set of relationships such as for different types of whole-part relations, and would reflect the many constraints that apply to the relationships and properties.

Any metadata that is intended to be used for findability and reusability needs to recognize and accommodate the distinction between conceptual and operational ontologies as well as distinctions among different styles of operational ontology. Otherwise an ontology developed to define real-world meanings may get re-used in an ontology application for data processing, with generally bad results. The intended use of an ontology is an example of an ontological commitment. In a similar vein, metadata should be included for other ontological commitments, such as realism and 4D "extensional" ontologies[9].

4. Scoping the Community Role in Quality Ontologies

As we have seen there is a need for richly developed vocabularies to describe a given domain. Perhaps no domain has systematically developed findable ontologies to support data use more than the biomedical field (Smith et al., 2008). A good biomedical community example is what has been developed as part of the OBO Foundry (OBO, n.d.). Besides a large user base and consistent management activity, the OBO Foundry Tools feature operational rules that are easy to find and access, along with a suite of automated tools to validate rule compliance and a dashboard for easy understanding on an ontology's degree of conformity with each principle (OBO Foundry Tools, n.d.; Jackson et al., 2021). Taken as a whole this work illustrates:

- principles and practices to ensure that relevant biomedical ontologies are findable.
- the use of a modular approach to support alignments and extensibility, and
- systematic axiomatizations of concise definitions that users may use for search.

When properly executed by a sizable, federated community, these principles can help improve overall quality and interoperability between ontologies

---

[9] A 4D ontology regards individual objects as being four-dimensional, having spatial and temporal parts, and existing immutably in space-time.

whose use in turn supports the goals of FAIRness.

Starting with the first principle above, one may find relevant conceptualized biomedical and associated domain vocabularies to make data FAIRer just by using the OBO registry (OBO Registry, n.d.). Here one finds persistently available, versioned copies of ontologies represented in OWL that are accessible using standardized interfaces. Protocols for retrieving well managed ontologies from repositories like BioPortal and the OBO Foundry are explicit and easy for both humans and machines. Furthermore, storage of ontologies on GitHub repositories include well-defined mechanisms to obtain authorization for access to versioned ontologies.

Of great importance is finding information about an ontology's coverage and scope. This essentially concerns the relevance of an ontology and what it claims to cover (Matentzoglu et al., 2018). In practice, as noted earlier, finding a relevant and accessible ontology requires good searching skills and experience followed by consideration of what is known about the ontology. General guidance for creating rich metadata, in the form of meta-ontologies, and registering them in trustworthy repositories can also help with other FAIR principles such as accessibility. But, as we have seen, there are a wide range of approaches to metadata for ontologies. Accordingly we should add richer information such as taxonomic classes and property relations, which may facilitate data integration and interoperability, and whether a state-of-the art, general ontological methodology or framework has been used[10].

A good illustration of such considerations in the biomedical domain was driven by the COVID-19 pandemic. This pandemic challenged healthcare systems and research worldwide as worldwide data was collected, but needed to be integrated and made broadly and rapidly available to researchers. Unprecedented amounts of data sourced from public health surveillance were put to varied uses. These included vast amounts of real-time monitoring of outbreaks, the analysis of current and forecast trends, and analysis of news reports and organizational briefings about utilization. Numerous heterogeneous information systems were used by institutional researchers and hospitals capturing virological, epidemiological, and clinical characteristics. This resulted in fragmentation and the creation of multiple interoperable data "silos."

---

[10] For example, the OBO Foundry (Jackson, Rebecca et al., 2021).

Rapid identification of a data integration problem quickly led to standardization efforts, starting with data about infections. In the OBO Foundry the Coronavirus Infection was formally defined using concepts from an aligned suite of modular interoperable ontologies. These include the Infectious Disease Ontology (IDO) and the Chemical Entities of Biological Interest (ChEBI), as well as the Human Phenotype Ontology for human host phenotypes. Updated standardization for COVID-19 was guided and simplified by the use of semantic resources and incorporation of knowledge organization systems (taxonomies and vocabularies as well as OBO ontologies) as background knowledge. This started with designing a common space that included a standard vocabulary defining the symptoms of a "Coronavirus Infection:"

> is characterized by fever, cough and shortness of breath and that has a material basis in SARS-CoV-2.

This definition can easily be found in trusted places, based on sound principles and practices like the OBO-Foundry. It can also be found using special ontology search engines such as the Ontology Look up Service (OLS) or the European Bioinformatics Institute (EBI-EMBL). The standard definition found in these sources includes a common identifier that is not only in ontologies and Wikipedia, but is axiomatized throughout the OBO Foundry. Using OLS we find the same (unambiguous) concepts cross referenced in different access tools. Along with other material it can be found in the following:

url:https://www.cdc.gov/coronavirus/2019-ncov/about/index.html
url:https://www.ncbi.nlm.nih.gov/pubmed/?term=32007145
url:https://www.ncbi.nlm.quadih.gov/Taxonomy/Browser/wwwtax.cgi?id=2697049

To flesh out the COVID-19 model some vocabularies of the core IDO model were enhanced by introducing the metadata/ontology vocabulary item *acellular*. This captures the idea of a structure to cover viruses along with *other acellular entities* that are studied as part of virology. These allowed distinguishing between *infectious agents* (i.e., organisms with an *infectious disposition*) and infectious structures (i.e., acellular structures that have an infectious disposition). Several new extensions in IDO virus ontology (VIDO) (Babcock et al., 2021) and the Coronavirus Infectious Disease Ontology (CIDO) (He et al., 20210) were developed and aligned with part of IDO so these ontology modules shared the same vocabulary (e.g., the concept of *Symptom*). Together these were designed to align and provide broad coverage across various aspects of the infectious disease domain such as allowing simple extensions to new pathogen-specific ontologies, which helped identify drug candidates that could be repurposed for an effective and safe COVID-19 treatment. Over 90 chemical drugs and antibodies against human coronavirus diseases were identified early on by mapping anti-coronavirus drugs to ontology identifiers from ChEBI and drug data using semantic similarity analysis (Liu et al., 2020).

This work shows the role of extendable modular domain ontologies working together to support interoperability at both the data and ontological level. The common OBO metadata also support other semantic resources like the creation of KGs which put diverse data together. It also illustrates the incremental way that modular ontologies may be matured over time (Berg-Cross, 2022).

5. Conclusion and Challenges for Grounding Metadata

We have seen some challenges pursuing shallow, objective grounding of concepts for approaches like FAIR maturity indicators. In the case of data these rely on too simple and shallow a semantic view of operationally defined meanings. In conclusion we can understand some of the problems of meaning that arise because of the recursive design considerations for the FAIRness evaluation framework. This reflects the insistence that all components of the evaluation framework should themselves be FAIR and would apply across the full range of digital forms of semantic resources. Most notably they apply to controlled vocabularies and taxonomies used to play the role of metadata along with ontologies that also play an important role in making metadata FAIR. These all need to be FAIR. However, generating a semantic model is often the most time-consuming step of what goes into systematically making data and metadata FAIR (Jacobsen et al., 2020). Moreover, users have to find and access a relevant ontology first to support this type of interoperability. But domain ontologies typically don't fully cover domain user interest or provide meta-ontology links to connect relevant ontologies; and, as we have argued, there is often the issue of alternate ontologies as well as other forms of semantic resources that might provide additional, possible conflicting information. Although ontologists have developed several guiding principles for documenting ontologies, finding and developing the range of resources needed for proper scoping and

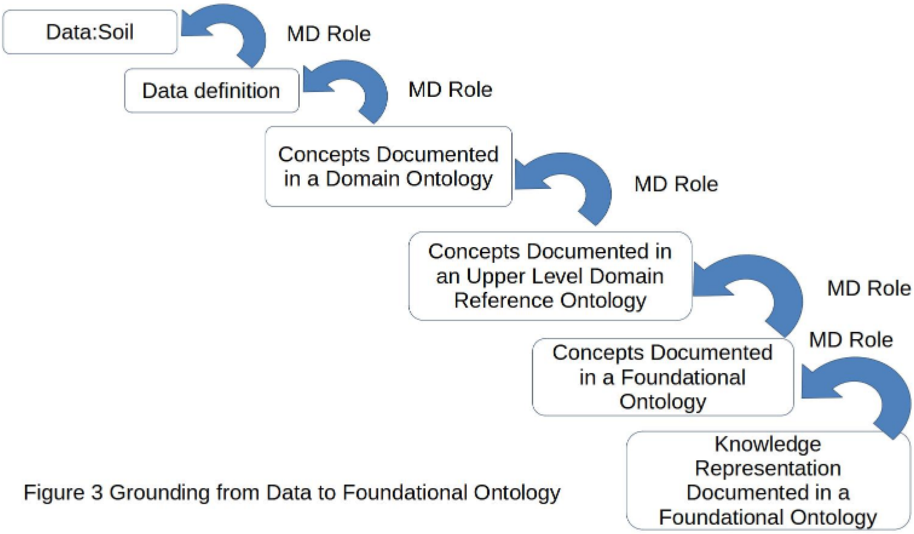


Figure 3 Grounding from Data to Foundational Ontology

coverage can be difficult for ordinary users. Grounding a whole suite of semantic resources using a meta ontology is still more difficult.

Metadata play a key role from finding to interoperation. An example is the diversity of conceptualizations of soil that was mentioned previously. What might be a simple example of rich metadata usable to start formally disambiguating many alternate definitions? One might think that the source for this richness is found in the related schemas of ontologies. However, these axiomatized schemas need to be grounded in community agreed upon conceptualizations rather than arbitrary semantic resources expressed in various ways. One can start with a domain ontology as shown in Figure 3, and one can point to an example of an axiomatized schema for soil in ENVO used as a previous example. The richness of an axiomatized definition relies in part on further definitions of terms and their axiomatizations supporting the soil definition. Thus "porous" is further defined in text as "a porosity quality inhering in a bearer by virtue of the bearer's being capable of admitting the passage of gas or liquid through pores or interstices." In this way rich metadata becomes more richly connected, replacing simple value pairs with a complex axiom pattern, which are, in turn, grounded in rationalized and useful definitions agreed upon by a community. There is a clear illustration of recursion of semantic grounding in such examples.

Having a common meaning is central to grounding, which implies that we need some better requirements for the grounding of meta ontologies. The concept of "porosity", for example, is not axiomatized in ENVO (2002). But some upper domain reference models do offer a range of basic concepts that might suffice for such grounding. These use basic concepts of containers and voids (Brodaric and Hahmann, 2014; Brodaric, Hahmann, and Gruninger, 2019). As shown in Figure 3 this takes us one step further in grounding, and this can be further grounded in something like the Unified Foundational Ontology (UFO) (Guizzardi et al., 2022). The UFO combines theories from philosophical reasoning about formal ontology with cognitive science and linguistic theory. Operationally it is composed of a handful of micro-theories that cover fundamental conceptual modeling notions such as kinds, primitive types and collectives. Still more grounding can be obtained by including concepts to describe the type of class a "kind" is and to detail meta-relationship (logical) types like "reflexivity" and "symmetry".

Some decisions on grounding may be found by studying conversations within a community and/or by a fiat decision that can be defended practically. In other cases physical observations may be used, and objective rules may be established.

Some ontology management activities may be needed as a result of search activities. In practice, when a simple search for terms of interest within an identified, relevant ontology is unsuccessful, new ontology terms are often defined and added to the existing ontologies. But this must be done cooperatively so that related ontologies and their users are aware of the changes. If the structure of an existing ontology seems unsuitable, new ontologies may have to be developed. An example of this is the existing mismatch between the SWEET and ENVO ontologies, which were independently developed and which have very different structures (Pouchard and Huhns, 2014; Karam at al, 2020). However, this is a time-consuming process that often needs to be undertaken with a team of domain experts in consultation with ontology experts.

References

Amdouni, Emna, Syphax Bouazzouni, and Clement Jonquet. (2022). "O'FAIRe makes you an offer: Metadata-based Automatic FAIRness Assessment for Ontologies and Semantic Resources."

Aminu, E. F., Oyefolahan, I. O., Abdullahi, M. B., and Salaudeen, M. T. (2020). A review on ontology development methodologies for developing ontological knowledge representation systems for various domains.

Babcock, Shane et al. (2021). “The infectious disease ontology in the age of COVID-19.” Journal of biomedical semantics 12.1: 1–20.

Baclawski, K., Bennett, M., Berg-Cross, G., Schneider, T., Sharma, R., Singer, J., and Sriram, R. D. (2021). Ontology Summit 2020 Communiqué: Knowledge Graphs. *Applied Ontology*, 16(2), 229–247.

Berg-Cross et al. (2021). Ontology Summit 2021 Advances in Semantic Harmonization. https://bit.ly/3pnkwdG

Berg-Cross, G. (2021). Issues in Incrementally Adding Better Semantics to Knowledge Graphs. J. Wash. Acad. Sci. 107(3): 33–58.

Berners-Lee, Tim, and James Hendler. 2001). “Publishing on the semantic web.” Nature 410.6832: 1023–1024

Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3.

Blomqvist, Eva, Torsten Hahmann, and Karl Hammar, eds. (2021). Advances in Pattern-based Ontology Engineering. Vol. 51. IOS Press.

Brodaric, Boyan, and Torsten Hahmann. (2014). “Toward a foundational hydro ontology for water data interoperability.”

Brodaric, Boyan, Torsten Hahmann, and Michael Gruninger. (2019). “Water features and their parts.” Applied ontology 14.1: 1–42.

Cimmino, Andrea, Alba Fernández-Izquierdo, and Raúl García-Castro. (2020). “Astrea: automatic generation of SHACL shapes from ontologies.” European Semantic Web Conference. Springer, Cham.

Crosswell, David. (n.d.) “Why your data must be FAIR”, https://bit.ly/3GF32UN accessed 12/12/2022

De Smedt, Koenraad, Dimitris Koureas, and Peter Wittenburg. (2020). “FAIR digital objects for science: From data pieces to actionable knowledge units.” Publications 8.2: 21.

Devaraju A, Mokrane M, Cepinskas L, Huber R, Herterich P, Vries J, Diepenbroek M. (2021). From conceptualization to implementation: FAIR assessment of research data objects. Data Science Journal, 20 (1): 1–14.

Disease Database. (2022) The Disease Database of 4000 Diseases, Disorders and Illnesses. http://bit.ly/3CQYoBX

Dutta, Biswanath, Durgesh Nandini, and Gautam Kishore Shahi. (2015). "MOD: metadata for ontology description and publication." International Conference on Dublin Core and Metadata Applications.

ENVO: Environmental Ontology. (n.d.). http://environmentontology.org/

FAIR Maturity Indicator. (2019). Gen2-MI-I1B. http://bit.ly/3Wc2oUv

Garijo, Daniel, Oscar Corcho, and María Poveda-Villalón. (2921). "FOOPS!: An Ontology Pitfall Scanner for the FAIR principles." Intern. Sem. Web Conf. (Posters/Demos/Industry).

Gruninger, M., Bodenreider, O., Olken, F., Obrst, L. and Yim, P. (2008). Ontology Summit 2007 – Ontology, taxonomy, folksonomy: understanding the distinctions. Applied Ontology, 3(3), 191–200.

Guizzardi. G. (2020). Ontology, Ontologies and the "I" of FAIR. Data Intelligence, 2(1-2):181–191.

Guizzardi, G., Botti Benevides, A., Fonseca, C. M., Porello, D., Almeida, J. P. A., and Prince Sales, T. (2022). UFO: Unified foundational ontology. Applied ontology, 1–44.

Hartmann, Jens et al. (2005). "Ontology metadata vocabulary and applications," OTM Confederated International Conferences "On the Move to Meaningful Internet Systems." Springer, Berlin, Heidelberg.

He, Y., Yu, H., Ong, E., Wang, Y., Liu, Y., Huffman, A., ... and Smith, B. (2021). CIDO: The community-based infectious disease ontology.

Heflin, J., and Hendler, J. (2000). Semantic interoperability on the Web. https://bit.ly/3pcuhkm

Jackson, Rebecca et al. "OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies." Database 2021.

Jacobsen, Annika et al. (2020). "A generic workflow for the data FAIRification process." Data Intelligence 2.1-2 (2020): 56–65.

Janowicz, K., Hitzler, P., Adams, B., Kolas, D., Vardeman, I. et al. (2014). Five stars of linked data vocabulary use. Semantic Web 5(3), 173–176.

Karam, Naouel et al. (2020). "Matching biodiversity and ecology ontologies: challenges and evaluation results." The Knowledge Engineering Review 35.

Liu, Y., Chan, W.K., Wang, Z., Hur, J., Xie, J., Yu, H., He, Y. (2020). Ontological and Bioinformatic Analysis of Anti-Coronavirus Drugs and Their Implication for Drug Repurposing against COVID-19. Preprints.org 2020, 2020030413. https://bit.ly/3HS8hSi

Maron, D. and Feinberg, M. (2018). “What does it mean to adopt a metadata standard? A case study of Omeka and the Dublin Core”, Journal of Documentation, Vol. 74 No. 4, pp. 674–691. http://bit.ly/3H6I2b1

Matentzoglu, Nicolas et al. (2018). “MIRO: guidelines for minimum information for the reporting of an ontology.” Journal of biomedical semantics 9.1: 1–13.

Mazimwe, Allan, Imed Hammouda, and Anthony Gidudu. (2021). “Implementation of FAIR Principles for Ontologies in the Disaster Domain: A Systematic Literature Review.” ISPRS International Journal of Geo-Information 10.5: 324.

Mikroyannidi E, Iannone L, Stevens R, Rector AL. (2011). Inspecting regularities in ontology design using clustering. In: The Semantic Web - ISWC 2011 - 10th International Semantic Web Conference, October 23–27, 2011, Proceedings, Part I. Bonn, p. 438–53. https://bit.ly/3CO2ISD

OBO Foundry. (n.d.). Home page http://bit.ly/3HfTOQx, accessed 12/23/2022

OBO Foundry Tools. (n.d.). http://bit.ly/3XzEtPI, accessed 12/23/2022

OBO Relation Ontology. (n.d.). http://bit.ly/3XviHg3, accessed 12/22/2022

OBO Registry. (n.d.). https://bit.ly/3XwigSR, accessed 12/23/2022

ODP. (2022). https://bit.ly/3IK0gAi

Palma, R. et al. (2020). Global Soil Information System (GloSIS) Ontology. Tech. Rep., SIEUSOIL project, https://bit.ly/3IPZ52k (last access: 1 November 2022).

Pandit, Harshvardhan J., Declan O’Sullivan, and Dave Lewis. (2018). “Using Ontology Design Patterns To Define SHACL Shapes.” WOP@ ISWC.

Pouchard, L. C., Huhns, M. N. (2014). RDF mappings between SWEET and ENVO ontologies. Purdue University Research Repository. doi:10.4231/R7BZ63ZG

Pourabdollah, Amir et al. (2012). “Towards a standard for soil and terrain data exchange: SoTerML.” Computers and Geosciences 45: 270–283.

Poveda-Villalón, María et al. (2020). “Coming to terms with FAIR ontologies.” Intern. Conf. on Know. Eng. and Know. Management. Springer, Cham.

RDA (2020). FAIR Data Maturity Model Working Group. FAIR Data Maturity Model: specification and guidelines. Research Data Alliance. DOI: 10.15497/RDA00050.

Smith, B. et al. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat. Biotechnol. 25: 1251–1255.

The Ontology Lookup Service. (2022). http://bit.ly/3XetyeD, accessed Dec. 9, 2022.

Todd-Brown, Kathe and Berg-Cross, Gary. (2022). Laying the bridge between soil data, knowledge, and semantics, July 2022 ESIP Meeting, http://bit.ly/3XeEVn1

Toulet, Anne, Biswanath Dutta, and Clement Jonquet. (2018). Assessing the practice of ontology metadata: A survey result. Diss. University of Montpellier.

Wilkinson, Mark D. et al. (2016). "The FAIR Guiding Principles for scientific data management and stewardship." Scientific data 3.1: 1–9.

Wilkinson, Mark D. et al. (2018). "A design framework and exemplar metrics for FAIRness." Scientific data 5.1: 1–4.

GARY BERG-CROSS is a cognitive psychologist (PhD, SUNY-Stony Brook) whose professional life included teaching and R&D in applied data and knowledge engineering, collaboration, and AI research. A board member of the Ontolog Forum he co-chaired the Research Data Alliance workgroup on Data Foundations and Terminology. Major thrusts of his work include reusable knowledge, vocabularies, and semantic interoperability achieved through semantic analysis, formalization, capture in knowledge tools, and access through repositories.

SAGE ARBOR has worked across a broad biomedical spectrum from drug design to clinical trial analysis including: project management of developers/analysts of globally distributed labs (PMP certified), electronic medical record datamining (SQL & noSQL), python/pandas coding, data segmentation, 6σ improvement, pathway mapping, and computational drug design & synthesis. He has worked in academia and multiple companies (e.g., Pfizer & Dupont) using data-mining terabytes to increase the transparency of information and efficacy of solutions. Since joining DCRI in 2021 he has worked to aggregate all NIH funded COVID data in the RADxUP project and led the agile team developing the web portal to upload, QC, and display aggregated data. He holds an Ph.D. in biochemistry from Washington University in St. Louis, MO, and a B.S. in biology and chemistry from Duke University.