Improving Biomedical Data by Improving the Information Artifact Ontology

Eric C. Merrell [1,2] and Robert M. Kelly [2,3]

[1] Department of Philosophy, University at Buffalo, 135 Park Hall, Buffalo NY 14260, USA
[2] National Center for Ontological Research, University at Buffalo, Buffalo NY, USA
[3] Romanell Center for Clinical Ethics and Philosophy of Medicine, University at Buffalo, Buffalo, NY

Abstract

The Information Artifact Ontology (IAO) is an ontology developed to tag data about information entities, especially ones relevant to the biomedical domain. IAO includes the class **information content entity (ICE)**, which is intended to capture all information entities. However, the direct subclasses (or 'children') of **ICE** (the 'parent' class) are distinguished in conflicting ways, which disposes extensions of IAO toward asserted multiple inheritance. This, in turn, breaks the well-known asserted single-inheritance rule of best practice for ontology building. We term the kinds of distinctions used to divide **ICE** into subclasses 'cross-cutting distinctions'. We offer a proposal for resolving IAO's use of cross-cutting distinctions in a way that retains the classes in IAO while adhering to asserted single inheritance and the principles of ontological realism.

Keywords

Asserted Inheritance, IAO, Inferred Inheritance, Interoperability, Multiple Inheritance, Ontological Realism

1. Introduction

The Information Artifact Ontology (IAO) was developed by the Open Biological and Biomedical Ontology (OBO) Foundry [8] to tag data about information entities. In this paper, we identify an ontological problem for effectively extending IAO into domain-level ontologies that adhere to realist principles of ontology building underlying Basic Formal Ontology (BFO) and adopted by the OBO Foundry. This problem stems from IAO's top-level class hierarchy. Modifying IAO's class hierarchy in the way we propose will enhance the interoperability of IAO, allowing data about information entities to be shared across domains. We explain this problem in detail below and propose a solution consistent with the principles of ontological realism [4]. Due to space limitations, we must refer the reader to some existing defenses of the realist principles underlying BFO and adopted by the OBO Foundry [3][4][8].

2. Background

One motivation for an information ontology is to accurately record and interpret patient healthcare data. This challenge becomes especially salient when those data are shared between healthcare providers at different organizations. Consider a general practitioner that sends records to a specialist to whom they have referred a patient, or cases where a patient must be transferred from one hospital to another. There is no guarantee (nor is it likely) that medical institutions collect, organize, label, or store patient information in the same way. As a result, it may be unclear to a recipient how to understand shared information. Detailed examples of this are discussed by Werner Ceusters and Barry Smith [1][2], and their proposed solution is to develop a referent tracking system (RTS) that uses a realist ontology. The benefit of an RTS is that it reduces ambiguity of data across organizations and domains. An information ontology like IAO is crucial in an RTS because each instance of any type of entity

International Conference on Biomedical Ontologies 2021, September 16–18, 2021, Bozen-Bolzano, Italy
EMAIL: ericmerr@buffalo.edu (A. 1); rkelly2@buffalo.edu (A. 2)
ORCID: 0000-0002-0193-3174 (A. 1); 0000-0002-6847-4726 (A. 2)
© 2021 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)
CEUR-WS.org/Vol-3073/paper9.pdf

(including information entities) must be assigned its own unique resource identifier (URI), a piece of information, to be represented in the ontology correctly.

While one aim of an RTS and of ontologies more generally is to reduce the ambiguity in data, ambiguity can still be introduced if the ontology allows classes to have more than one asserted parent (called 'asserted multiple inheritance') because cross-cutting distinctions allow for lower-level classes to reasonably be subsumed under more than one parent class. This breaks the principle of best practice used in realist ontologies known as 'asserted single inheritance' [3]. Unfortunately, IAO often uses cross-cutting distinctions when asserting subtypes of **information content entity (ICE)**. While some debate the appropriate use of single versus multiple inheritance [14], we follow Rector [13], Mungall [14], and Arp, Smith, and Spear [3] in maintaining that multiple inheritance should never be asserted in an ontology, but should only be inferred from its axioms and class restrictions.

IAO itself, it should be emphasized, does *not* suffer from asserted multiple inheritance. However, the use of cross-cutting distinctions within its upper-level class hierarchy means either that (i) asserted multiple inheritance will be needed to extend IAO, or (ii) IAO can be extended in multiple, contradictory ways. Both can lead to the extensions of IAO being incompatible.

3. The Problem of Cross-Cutting Distinctions

Cross-cutting distinctions occur when entity types are subdivided in such a way that at least some of the new subtypes can consistently be subclasses of two or more super-classes in the primitive class hierarchy. We assume here that universals (genuine types in reality) do not crosscut. As a quick example, consider that no genuine subtype of **continuant** will also be a genuine subtype of **occurrent**. We assume the same is true for any genuine types and subtypes. Cross-cutting distinctions correspond to violations of Rector's second principle of ontology building, namely that "the principle of specialization should be subsumption (as opposed to, for example partonomy) and should be based on the same, or progressively narrower criteria, throughout" [13]. There is nothing inherently inaccurate (as far as accurately representing reality is concerned) with making cross-cutting distinctions. Still, when constructing ontologies, it can lead to practical inaccuracies in tagging data.

Consider a simple example. Suppose we have a food ontology including the class **fruit**, which we divide into four subclasses using two distinct criteria: **citrus fruit** and **non-citrus fruit**, on the one hand, and **orange-colored fruit** and **non-orange-colored fruit,** on the other. The difficulty is that apricots are orange non-citrus fruits, while limes are green citrus fruits. Thus, each has two categories under which they could fall. See Figure 1 below for an illustration of this point.

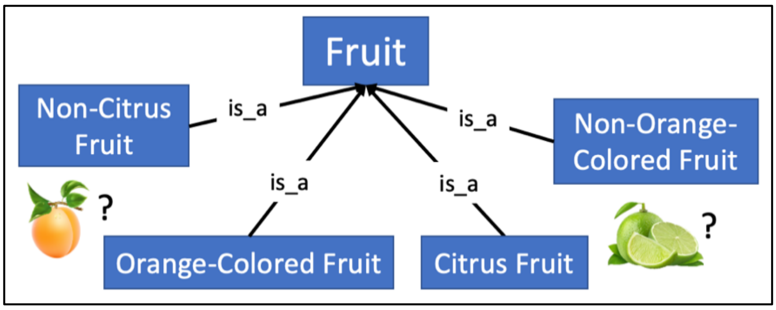

Figure 1: Potential for Asserted Multiple Inheritance

Contrast this with a division of **fruit** that keeps **citrus fruit** and **non-citrus fruit**, but that further divides these subclasses into **orange citrus fruit**, **non-orange citrus fruit**, **orange non-citrus fruit**, and **non-orange non-citrus fruit**. These divisions might seem even more gerrymandered, but they would follow Rector's second principle above since they take a single criterion (e.g., citrus vs. non-citrus) and start to make it progressively narrower (e.g., orange citrus vs. non-orange citrus). This would lead to only one classification for apricots and only one for limes, preserving asserted single inheritance.

Now, IAO includes several subclasses of **ICE**, including **document**, **document part**, **textual entity**, **figure**, and **directive information entity**, each of which is relevant to assert in certain contexts. It would not be a problem to assert these classes, so long as the correct equivalency axioms and class restrictions are associated with them, because this allows ontology reasoners to *infer* multiple inheritance from the axioms and restrictions given. However, because different criteria are used to identify the different subtypes of **ICE** (and so the distinctions crosscut), and because the required equivalency axioms and restrictions are not always given, subclasses (and sub-subclasses, and so on) further down the **ICE** branch might belong to more than one primitive child of **ICE**. This ultimately hinders the ability to extend the ontology using asserted single inheritance [4][9][13].

3.1. Cross-cutting and Reports

IAO uses at least three different criteria to distinguish subtypes of **ICE**: (1) the *content* of an ICE; (2) the *structure* of an ICE; and (3) the *purpose* of an ICE. The issue is that a single type of ICE can have a structure indicating it should be classified as one sort of subtype of **ICE** but content that would indicate classifying it as another sort. A good example is the IAO class **report**.

> **IAO:Report** =def. A document assembled by an author for the purpose of providing information for the audience. A report is the output of a documenting process and has the objective to be consumed by a specific audience. Topic of the report is on something that has completed. A report is not a single figure. [5]

According to this definition, a report is a document. From IAO's definition of 'document', it follows that a report is also a "collection of information content entities" [5]. Examples of documents given in IAO include journal articles, lab notebooks, and books. Accordingly, the class **journal article** was made a subclass of **report**. However, because journal articles and grant proposals, among other kinds of reports, (usually) consist primarily of propositions, the class **journal article** is also suited to be a subclass of **narrative object**, which IAO defines as an ICE that is a set of propositions. Intuitively, **report** (and by extension, **journal article**) ought to be a subclass of both **document** and **narrative object**. Yet, it is impossible to *infer* this multiple inheritance using the axioms present in the ontology.

From within the framework of ontological realism, this indicates (though does not entail) that **document** and **narrative object** are both universal classes, since defined classes (or 'collections of particulars' to use Smith and Ceusters terminology [4]) ought to have logical restrictions associated with them. Thus, defined classes or collections of particulars in the context of a realist ontology can be thought of as describing entities that match a certain pattern of universals. They are not defined classes in the sense used by Rector [13] because they are not defined using a set of *necessary and sufficient* conditions, but only *necessary* conditions, which makes them a primitive class on Rector's view. This does not mean that one class or other need be eliminated from IAO. It simply means that a single metric needs to be used to create the basic ontological hierarchy of IAO. More work would be needed to define classes that do not use that metric in terms of other universal classes that form IAO's backbone. This would avoid asserted multiple inheritance and enable inferred multiple inheritance [3][13][14].

3.2. Cross-cutting and Directive ICEs

A similar cross-cutting problem arises with directive ICEs. A directive ICE is any ICE that indicates to an agent which course of action to take (such as a recipe). Many directive ICEs are also documents. An email instructing someone to perform a certain task is a document *and* a directive ICE. Similarly, a to-do list for an employee would simultaneously be both a directive ICE and a document. Each of these would also be narrative objects since they are all composed of propositions (though we do not hold that all directive ICEs *must* be narrative objects). Many kinds of ICE besides **document** and **narrative object** are simultaneously documents and directive ICEs.

The problem is that making both **directive ICE** and **document** sibling classes under **ICE** without making one a defined class requires either that (i) the ontology include asserted multiple inheritance, or (ii) subclasses of **directive ICE** cannot also be subclasses of **document** or **narrative object**. The former breaks the principle of asserted single inheritance, while the latter steers us away from ontological realism since, as noted, it is clear that some directive ICEs are also documents and narrative objects. We acknowledge that OWL ontologies are built under the open world assumption [9][3], and so failing to assert that a subclass of **directive ICE** is also a subclass of, say, **document**, does not entail that the absent assertion is false. Still, IAO's current structure *constrains* our ability to actually assert what is true, at least if we intend to avoid asserted multiple inheritance. This is what we mean by the second disjunct above steering us away from realism. To illustrate this, consider IAO's **plan specification**.

> **IAO:PlanSpecification** =def. A directive information entity with action specifications and objective specifications as parts that, when concretized, is realized in a process in which the bearer tries to achieve the objectives by taking the actions specified. [5]

Thus, **plan specification** is a subclass of **directive ICE**. Supposing we want to extend **plan specification** with the class **cooking recipe**, we run into a problem. Instances of **cooking recipe** (we might assume) consist of any plan specification that explains how to cook some food item. This would include cookbooks, web pages on food preparation, and even old, hand-written family recipes. Each of these is a document, meaning that including the most amount of ontological information about these recipes requires **cooking recipe** to be a subclass of **document**. But this cannot be done within the current class hierarchy of IAO without violating asserted single inheritance.

3.3. Cross-cutting and Document Parts

The IAO class **document part** is extended with a variety of subtypes. Many of these, however, also fit the definition for other *sibling* classes of **document part**. Once again, this can result in either asserted multiple inheritance or departure from ontological realism if the ontology is extended. First, consider the definition of the following subclass of **document part**.

> **IAO:MethodsSection** = def. A part of a publication about an investigation that is about the study design of the investigation. [5]

This class extends **document part**, but it could also reasonably extend the class **textual entity** because a methods section is a collection of markings and symbols intended to be understood as words. Thus, there are at least two potential parent classes of **methods section**. Again, this results from the lack of a uniform method for asserting subclasses of **ICE**.

Next, consider the definition of the following additional subclass of **document part**.

> **IAO:FiguresSection** = def. A part of a document that contains one or more figures. [5]

An instance of **figures section** may itself consist of but a single figure. This means that instances of **figures section** might be identical to one or more figures, or they might be any part of a document that contains a figure, such as a methods section. Accordingly, defining **figures section** in this way could lead to asserted multiple inheritance in the ontology if **figures section** or **methods section** have asserted subclasses. This would not be a problem, if **figures section** were a defined class with associated logical restrictions, which would allow inferred multiple inheritance.

4. Resolving Cross-cutting Distinctions

The prevalence of cross-cutting distinctions in IAO suggests that the difference between defined and primitive classes is not properly acknowledged. Extending IAO is thereby prone to either a hierarchy including asserted multiple inheritance or a departure from ontological realism. Abiding by ontological realism means asserting classes based on the way the world actually is [3]. Practically, this means ensuring that the primitive class hierarchy consists only of what are taken to be universals. When cross-cutting distinctions are present in a primitive class hierarchy of universals (in a realist ontology), as in IAO, it indicates that careful attention is not being paid to establishing an asserted class hierarchy of only universals – classes that correspond to genuine types in reality [3]. This is not to say that cross-cutting distinctions cannot accurately describe reality (e.g., there really are orange fruits and non-citrus fruits), only that asserting such distinctions in a primitive class hierarchy should be avoided. We propose that by establishing a primitive hierarchy of universals, we can abide by both realist principles and the principles established by Rector [13], which would allow IAO to be more effectively extended.

For example, it seems obvious at first blush that documents have parts (the document you are reading now is divided into many sections, after all), and so it makes intuitive sense to assert **document part** as a class in an information ontology. However, document parts do not seem to be *genuinely different kinds* of entities in virtue of being parts of documents. The same way that a human being does not become a different *kind* of thing when they acquire a new job with a new set of duties and responsibilities. The different parts of the documents are real, to be sure, but it does not follow that they are themselves genuinely different types of things. (Consider that the left half of your scalp is real but **left half of scalp** is a fiat part of an object and not a genuinely distinct class of entity.)

Differences in writing styles across fields entail that document parts, like introductions, can vary widely in their role and content. For instance, a methods and a figures section, or a figures section and an introduction, may overlap according to IAO depending on the content of that section. Additionally, a single part of a document might have multiple purposes, which would need to be distinguished in the ontology. Authors seem to write each document part for a particular purpose, and so we should distinguish the various functions that document parts (or even whole documents) can have. This would allow us to account for when a single part of a document subsumes the purposes of multiple extensions of **document part**. This suggestion follows Rector's second criterion of normalization, which uses the same principle of specialization when building the primitive backbone hierarchy of an ontology [13].

Accordingly, we propose distinguishing ICEs based on the *purpose* of those ICEs. This principle of differentiation facilitates constructing a consistent and effective information ontology since all ICEs are the products of some intentional creation process. Applying this principle leads to three high-level direct subclasses of **ICE**, which mirror those in another widely used information ontology: the Common Core's Information Entity Ontology (IEO) [6]. We have modified the IEO classes and propose they be added to the top level of IAO (or that the IAO definition be changed in the case of **directive ICE**). Where possible, we make use of the term 'represents' as distinguished from Ceusters' use of 'is about' [7]. This avoids contradictions in tagging ICEs that are not about anything (for example, a wrong diagnosis of cancer). These three classes are intended to be the only three primitive direct subclasses of **ICE.** However, this does not preclude there being defined classes that extend **ICE** directly.

1. **Descriptive ICE** =def. an ICE that describes some entity it represents or would be represented if such an entity existed.
2. **Designative ICE** =def. an ICE that refers to or names some entity or would if such an entity existed.
3. **Directive ICE** =def. a descriptive ICE that prescribes some action.

4.1. Change in Class Structure

We make **directive ICE** a child of **descriptive ICE** because all directive ICEs are necessarily descriptive. We follow Barton and colleagues in maintaining that directive ICEs "represent how to perform some actions, and such actions are then performed on the basis of this description" [11]. For an ICE to prescribe an action, it must also *describe* the action that it is prescribing. Suppose a parent leaves a note for their daughter that reads, "Please cut the grass." This note carries a directive ICE only by describing the kind of action that is also prescribed (grass cutting). The description is necessary to differentiate which action the directive ICE prescribes.

We admit that conceiving of all directive ICEs as containing descriptions sets a low bar for an information entity being descriptive. A note reading "please cut the grass" does not describe very much. It contains no description of grass, of cutting, nor of what it would mean to cut grass specifically as opposed to cutting a steak or paper. There is also no way to tell which grass is being referred to simply by what is written on the note – context is needed to understand what is meant.

However, there is another sense in which the desired action is described in the note. In virtue of specifying that they want the grass cut, the parent has already implicitly described a process type. Using a knife to cut steak, or a pair of scissors to cut paper, would not satisfy the description of grass cutting. That the prescription in the note is not satisfied by cutting a steak implies that one's steak cutting does not match the *description* of the process in the note. Hence, the note is, in part, descriptive. Moreover, as Barton and colleagues note [11], descriptions of actions to be performed can be more or less specific. The more specific a set of instructions, the more thoroughly those prescribed actions must be described.

Here, one might wonder why **designative ICE** was not made a parent of **descriptive ICE**. After all, it seems as though all descriptive ICEs would need to designate the various entities they describe. This is certainly plausible, but even if this is true, it does not indicate that descriptive ICEs are themselves designative ICEs. It only indicates that descriptive ICEs contain designative ICEs as parts. We lack the space to spell out exactly how this parthood relation works, but other work has been specifically focused on the mereology of information entities [10]. Further work is needed to apply these findings here.

Lastly, considering our principle of differentiation for ICEs can shed light on why **designative ICE** is not a parent of **descriptive ICE.** The purpose of descriptive ICEs is to describe a state of affairs rather than to designate or name an entity, even if descriptions often require designating entities. This is dissimilar from the relationship between descriptive and directive ICEs because the purpose of the latter *is* to describe some action(s) that ought to be carried out. Thus, the purpose of directive ICEs is partly descriptive in nature, which justifies **directive ICE** being a subclass of **descriptive ICE**.

4.2. Additional Defined Classes

While every ICE will ultimately be an instance of **designative**, **directive**, or **descriptive ICE**, IAO can also retain many of the useful classes it asserts (such as **document**, **textual entity**, and **document part**) by asserting them as defined classes. In Table 1, we propose samples of axioms (class restrictions) that could be added to existing IAO classes to make them proper defined classes. The classes in this table are defined so that they are not included in the primitive class hierarchy, which removes cross-cutting distinctions from it. We are neutral about whether the defined classes here are universals. Still, since they are defined classes, whether they correspond to universals will not infringe upon ontological realism. The list is not exhaustive and further work would be needed to make the necessary changes that would eliminate all the cases of cross-cutting distinctions within the primitive class hierarchy of IAO. Still, we hope that this provides an example for future modifications to IAO.

Table 1:
Proposed Axioms to be Added to IAO Classes

Class	Natural Language Axiom	Formalized Axiom
IAO:Document	All Documents are ICEs and have as a part some ICE.	∀(x)[IAO:Document(x) ↔ (ICE(x) ∧ ∃(y)(has_part(x,y) ∧ ICE(y)))]
IAO:TextualEntity	All Textual Entities are concretized by some Glyph Quality.	∀(x)[TextualEntity(x) ↔ (ICE(x) ∧ ∃(y)(is_concretized_by(x,y) ∧ GlyphQuality(y)))]
IAO:Figure	All Figures are concretized as some Two-Dimensional Pattern.	∀(x)[Figure(x) ↔ (ICE(x) ∧ ∃(y)(is_concretized_as(x,y) ∧ 2DPattern(y)))]

5. Conclusion

In this paper, we identified and described the problem of cross-cutting distinctions within IAO. These distinctions make it difficult or impossible to extend IAO consistently, according to the principles of asserted single inheritance and ontological realism. For those adhering to these principles of ontology development, we proposed a modification to the top-level structure of IAO. Our proposal eliminates cross-cutting distinctions in IAO’s universal hierarchy, and it develops logical axioms for inferring poly-hierarchies where needed. While there is still ample work to be done, we hope that our proposal can be used to enhance the effectiveness of IAO.

6. Acknowledgements

We would like to thank David Limbaugh and Barry Smith for their feedback on this work, as well as four anonymous reviewers for the immensely helpful feedback they provided.

7. References

[1] W. Ceusters, B. Smith, Tracking referents in electronic health records, Studies in Health Technology and Informatics 116 (2005) 71-76. PMID: 16160238.

[2] W. Ceusters, B. Smith, Strategies for referent tracking in electronic health records, Journal of Biomedical Informatics 39.3 (2006) 362-378. doi: 10.1016/j.jbi.2005.08.002.

[3] R. Arp, B. Smith, A. D. Spear, Building ontologies with Basic Formal Ontology, MIT Press, Cambridge, MA, 2015.

[4] B. Smith, W. Ceusters, Ontological realism: A methodology for coordinated evolution of scientific ontologies, Applied Ontology 5.3-4 (2010) 139-188. doi: 10.3233/AO-2010-0079.

[5] The OBO Foundry, Information Artifact Ontology, 2021. URL: http://www.obofoundry.org/ontology/iao.html

[6] M. Jensen, Information Entity Ontology, 2021. URL: https://github.com/CommonCoreOntology/CommonCoreOntologies/blob/master/InformationEntityOntology.ttl.

[7] W. Ceusters, An Information Artifact Ontology perspective on data collections and associated representational artifacts, Studies in Health Technology and Informatics 180 (2012) 68-72. PMID: 22874154.

[8] B. Smith, M. Ashburner, C. Rosse, et al., The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration, Nature Biotechnology 25 (2007) 1251–1255. doi: https://doi.org/10.1038/nbt1346.

[9] D. Allemang, J.Hendler, The semantic web for the working ontologist, 2nd ed., Morgan Kaufmann Publishers, Elsevier, MA, 2011.

[10] A. Barton, F. Toyoshima, J. Ethier, Clinical documents and their parts, in: Proceedings of the 11th International Conference on Biomedical Ontologies (ICBO), CEUR-WS.org, Bolzano, Italy, September 17, 2020, pp. A1-A11.

[11] A. Barton, L. Vieu, J. Ethier, Directing Actions, in: Proceedings of the Workshop on Foundational Ontology (FOUST III), Joint Ontology Workshops (JOWO 2019), CEUR-WS.org, Graz, Austria, 2019, pp. 1-10.

[12] S. Seppälä, A. Ruttenberg, B. Smith, The functions of definitions in ontologies, Formal Ontology in Information Systems 283 (2016) 37-50. doi: 10.3233/978-1-61499-660-6-37.

[13] A. L. Rector, Modularisation of domain ontologies implemented in description logics and related formalisms including OWL, in: Proceedings of the 2nd International Conference on Knowledge Capture (K-CAP '03), Association for Computing Machinery, New York, NY, 2003, pp. 121–128.

[14] C. Mungall, Monkeying around with OWL: Musings on building and using ontologies, 2021. URL: https://douroucouli.wordpress.com/2019/05/10/.