Differences Among Cell-structure Ontologies: FMA, GO, & CCO

Alan Au, Xiang Li, & John H. Gennari, Ph.D.
Biomedical and Health Informatics, University of Washington, Seattle
{aau, mstrxl28, gennari}*@u.washington.edu*

Abstract

When different groups create models or ontologies of the same knowledge domain, this creates challenges for knowledge sharing. To identify these challenges, we compare cellular structure as modeled by the Foundational Model of Anatomy (FMA), the Gene Ontology (GO), and the Cell Component Ontology (CCO). These ontologies all model the physical anatomy of a cell, and we expected them to be similar in scope. However, we discovered that the actual differences among them are substantial. These differences represent variations based on theory-driven vs. emergent construction, as well as differences in how small application ontologies like the CCO are created from reference ontologies. In this paper, we provide a description and analysis of these differences. By studying differences in language, granularity, breadth of coverage, and model organization, we hope to gain a better understanding of how to map between related ontologies.

Introduction

Ontologies are gaining widespread use within the rapidly growing bio-informatics community. This community is also committed to the notion of *knowledge sharing*, where results and data are made available to the whole research community (usually after publication). An important example is the Gene Ontology (GO), which enables researchers to use a common terminology for gene product annotation. More recently, a number of researchers have worked to expand and improve the Gene Ontology, allowing researchers and intelligent systems to use GO annotations for advanced knowledge manipulation and inference [1].

Unfortunately, as ontologies become more expressive and more prevalent, it will become more important that we develop strategies to share knowledge *across* multiple ontologies and knowledge sources. To answer queries that draw on multiple ontologies, it is necessary to map or link terms across those ontologies. Often, two or more groups may independently develop ontologies that cover the same, similar, or overlapping knowledge. Naturally, one might like to combine these efforts and incorporate data and knowledge from all sources.

This scenario describes our situation with respect to information about cellular structure as expressed by three different ontologies: The Foundational Model of Anatomy (FMA) [2], the Gene Ontology (GO) [3], and the Cell Component Ontology (CCO) [4]. While they all contain knowledge about cellular structure, we will show that these three ontologies are actually quite different. In this paper, we characterize and quantify the differences among the subsets of the FMA, GO, and CCO that deal with cellular structure. Rather than suggesting that one ontology is better than another, we aim to understand how they differ. In the long term, we hope that our methods and resulting categories of differences will generalize across any pair of ontologies.

The Foundational Model of Anatomy, the Gene Ontology, and the Cell Component Ontology

The main purpose of the Foundational Model of Anatomy is to provide a baseline model of human anatomy upon which to conduct other research [2]. Constructed using the Protégé frame-based system [5], the FMA makes extensive use of hierarchical "is-a" relationships to classify and interrelate terms. As a whole, the FMA is a formal, theory-based representation of anatomical structure from the organism level down to the macromolecular level. Because our goal is to compare ontologies that describe cellular structure, we selected only the subset of the FMA dealing with the cell, cell part, and macromolecule.

The Gene Ontology is widely used by researchers in molecular and cell biology to annotate gene products with function, process, and cell component information [3]. Created in 1998, the GO was originally devised as a common terminology for research concerning model organisms such as Drosophila (Flybase) and Arabidopsis (TAIR). As a community-based resource, it has since grown to include more than 17,000 terms covering a broad range of biology, from bacteria to Homo sapiens. Because we are only interested in the cellular component section, we do not discuss the process or function portions of the GO.

The Cell Component Ontology (CCO) was created as part of the AraCyc project for use in annotating anatomical parts from the Arabidopsis plant [4]. In contrast to reference ontologies like the FMA, the CCO is an application ontology intended to support descriptions of pathway knowledge for Arabidopsis. As a result, terms in the CCO are selected almost exclusively from the GO. This reflects a pur-

poseful limitation of the CCO ontology scope. In addition, the AraCyc group classifies all of the terms into a more rigorous "is-a" hierarchy, adding terms where necessary to support this structure.

While the FMA, the GO cell component, and the CCO cover similar domains, there are notable differences among them. At a high level, it is important to consider basic philosophical differences in ontology development. The FMA is designed as a reference ontology: It is not designed for any specific application, but instead designed to be used by multiple types of applications and users.[2] Thus, it is built according to a set of rigorous modeling principles, so that each term is carefully defined, appropriately located, and linked to other terms in the ontology [6]. Although focused on Homo sapiens, the FMA provides a framework for modeling generic anatomy, and thus, some higher-level terms are meant for mammals or vertebrates [2].

In contrast, the GO models a canonical cell across multiple species, and is designed for a specific purpose—annotation of genomic research. As a consortium-controlled ontology, new GO terms are added whenever consortium members believe that those terms are important for annotation. In cases where a term is species specific, the GO uses a "sensu" tag to indicate the corresponding organism. As a model of a canonical cell, the GO excludes information specific to particular cell types, unlike the FMA, which models some cell types explicitly.

The CCO is an "application ontology"; it is designed with a more specific set of users and tasks in mind, namely, supporting pathway knowledge in Arabidopsis. Thus, it focuses on plant cell anatomy, and only at a particular level of depth and detail. Because of this variation in scope, and because the CCO is derived from the GO, our analysis first looks at FMA versus GO, and secondarily at how the CCO differs from its parent ontology (the GO).

Methods

Our analysis is based on a copy of the FMA from Jan., 2006, a version of the GO from Feb., 2006, and the CCO obtained in Feb., 2006. However, because the three ontologies are designed so differently, we do not treat the three equally. In particular, we focus primarily on the FMA vs. GO comparison. These two ontologies are similar in size, and they were developed independently. In contrast, the CCO is a much smaller application ontology derived primarily from the GO—the great majority of its terms are directly linked (via GO ID numbers) to the matching terms in the GO. In fact, because of this relationship, the analysis of CCO vs. FMA is redundant with GO vs. FMA, so we will not discuss that comparison further.

There are 1,172 terms in our selection of the FMA and 1,807 terms in the cellular component portion of the GO. The CCO contains 150 terms. Before beginning our comparison, we removed 115 GO terms which were annotated as "obsolete" and removed 10 duplicate terms, reducing our selection of the GO to 1,682 terms.

Our initial analysis of the FMA and GO consisted of checking for direct term matches and synonym matches. The FMA and the GO both include lists of "synonym" terms which represent the same concept. We started our comparison using only FMA preferred terms and GO preferred terms. In the results section, we categorize synonymy as one form of ontology difference.

Initial analysis involved direct string matching to find the exact overlap between the GO and the FMA. Since they both cover cellular structure, we assumed that identical terms shared common semantic meaning. Surprisingly, there were only 147 exact string matches between the FMA and the GO. Thus our initial comparison of preferred terms left 1,025 terms from the FMA and 1,535 terms from the GO that did not match. Once synonyms were taken into account, the number of unique terms dropped to 972 for the FMA and 1,479 for the GO.

For the GO vs. FMA comparison, our task was to analyze these 2,451 unmatched terms in order to understand why and how the two sets of terms were different. As shown in the results section, we looked for categories and clusters of differences. We did an extensive literature review to understand the biological basis behind each term. We also conducted interviews with the anatomists from the FMA project to obtain their expert input. In addition, we made extensive use of ontology visualization tools (for both the GO and the FMA) to assist us with our comprehension of the terms, both semantically and structurally.

Because of its small size and direct linkages to the GO, we were able to compare CCO vs. GO by hand. Below, we begin with our findings for the GO vs. FMA comparison, and then discuss the CCO vs. GO analysis in this context.

Results

Although the FMA and the GO purport to cover the same cellular structure, they are quite different. From our analysis, we define four fundamental categories of differences: language, depth of scope, breadth of scope, and ontological organization.

Language differences account for two main types of simple differences. The first type includes differences due to synonymy. These are terms which share the same conceptual meaning while using

significantly different terminology. As an example, "lamellar body" and "keratinosome" refer to the same biological concept. However, there is no simple way to correlate the two terms unless their semantic meanings are known. The second type of language difference encompasses minor linguistic variations. This includes different grammatical forms of the same word, as well as minor variations in word order (e.g., axoneme of flagellum vs. flagellar axoneme). Differences in this category can be handled by using a synonym mapping table. This table can be constructed manually, or it can be built using an automated system to account for these linguistic variations.

Differences in ***depth of scope*** represent new content that exists in only one of the ontologies. These terms have a corresponding term in the other ontology but at a different level of granularity. We characterize these differences into two types of relationships: "is-a" and "part-of". The "is-a" type of depth difference refers to shared terms, where one ontology provides sub-categories of that term. For example, delta-tubulin (FMA) is-a tubulin (GO), and cell-matrix junction (GO) is-a cell junction (FMA). Alternately, the "part-of" type of depth difference indicates that one ontology provides more partonomic descriptions of that term. For example, matrix of endosome (FMA) is part-of endosome (GO), and actin cap (GO) is part-of actin (FMA).

The ***breadth of scope*** category is similar to depth of scope; both categories include new concepts that do not exist in the other ontology. However, the breadth of scope characterizes terms from one ontology that are not related to any other term in the other ontology by either "is-a" or "part-of" relationships. We characterize those differences into three sub-types: subjective categorization, domain specificness, and novel concepts. Subjective categorization differences refer to terms that are defined by either the FMA or the GO in regards to their own perceptions of cellular structure. For example, endomembrane system (GO) and nuclear matrix proper (FMA) are concepts that rely on the anatomist's view of how cellular structure should be segmented.

The domain specificness sub-type describes another fundamental difference between the FMA and the GO, in that they are used for different research purposes, with specific domains of interest. Therefore, they both choose to represent only those concepts that are relevant to their respective usage and interests. For example, the GO has non-human terms such as "viral capsid", and the FMA has terms for specific human cell types such as "nucleus of cone cell". We have labeled the third sub-type of breadth difference simply as "novel concept". Novel concept represents those terms that are neither subjectively categorized nor domain specific. For example, complexes (GO) and amino acids (FMA) are concepts that do not exist in the other ontology. In these two cases, the exclusion is by design—the ontology designers have chosen different boundaries for the domain of "cellular structure".

Table 1. The distribution of terms across our four categories of ontological differences between the FMA and the GO.

	GO not in FMA (1479 terms)	FMA not in GO (972 terms)
Language	50 (4%)	58 (6%)
Scope - Depth	211 (14%)	229 (23%)
Scope - Breadth	1144 (77%)	629 (65%)
Ontological Organization	74 (5%)	56 (6%)

Finally, the FMA and the GO also employ different ***ontological organization*** to express their different views of cellular structure. While the FMA focuses on an accurate structural representation, the GO describes a canonical view of the generic cell. Accordingly, a small percentage of terms are reserved for purely organizational purposes. These concepts typically do not map between ontologies, as they reflect the organizational design of an ontology rather than the modeled content. Therefore, unless two groups share similar ideas about how to organize the subject matter, differences in representation cannot be reconciled automatically.

We sorted each of the 2,451 non-matching terms into one of the four categories as shown in Table 1. Our analysis shows the largest category of differences is breadth of scope, and the smallest category of differences is language variation. In addition, our results show that the FMA has a similar percentage of language and organizational differences as the GO. The FMA has a slightly larger percentage of depth of scope differences than the GO, but a smaller percentage of breadth of scope differences.

Using these categories of differences, we next examined the CCO and its relationship to the GO. As a GO-based application ontology, 90% of the CCO terms were taken from the GO, and so it largely avoids problems with language differences. Of the remaining 10%, most were depth of scope differences. This includes "sensu" variations for terms like plasma membrane and granularity for terms such as chloroplast thylakoid membrane, providing a greater level of depth than the GO. A small number of terms (such as "space") represented ontological organization differences and corresponded to terms in the FMA. In general, the overall intersection between the CCO and the FMA was low

due to differences in species representation (CCO was created for Arabidopsis).

While an application ontology would not be expected to include more breadth than a general ontology, it is interesting to look at the breadth of the terms contained in the CCO. As might be expected, the majority of terms fall in the middle of the GO hierarchy. This demonstrates that the CCO includes fewer high-level terms than the GO. The CCO also features a decreased branching factor near the bottom areas of the ontology as compared to the GO, again indicating a reduced breadth of coverage.

Discussion

We find the results in Table 1 surprising and interesting along a number of dimensions. First, recall that we selected subsets of the FMA and the GO so that they would both cover only knowledge about cell structure. Thus, we initially expected small amounts of differences in scope of knowledge. Given the very small number of exact string matches, our first guess was that language differences would account for a large portion of the differences. As Table 1 shows, this was not the case. Instead, there were very significant differences in both depth and breadth of scope, even within the relatively narrow domain of cellular structure.

Next, given that there are significant differences in depth, we might guess that one ontology would provide more details than the other. In particular, we guessed that because the GO is at the genomic level, it would have more low-level detail than the FMA, which includes gross anatomy. However, it turns out that neither ontology is more detailed than the other. Instead, they are complementary; they include more and less depth in different places. Our results also show that the depth differences are split about evenly between "is-a" and "part-of" depth differences for both ontologies.

Table 1 shows that the largest variation in our categorization is the breadth of scope difference. The large breadth disparity between the FMA and the GO is mainly due to four sorts of different types of knowledge. First, as part of its structural knowledge about the cell, the GO includes more than 670 "complexes", or combinations of proteins. Although the FMA anatomists agree that complexes should be included in their ontology (personal communication), it currently does not include this knowledge. In a similar fashion, the FMA includes more than 400 terms that are lipids, glycoproteins, and amino acids. Just as the FMA does not include complexes, the GO intentionally does not currently include any of these terms. We categorize both of these breadth differences into the "novel concepts" sub-type.

Third, the FMA includes about 140 terms that have to do with particular cell types within the human (E.g., "cell body of olfactory receptor cell"). As mentioned earlier, because the GO represents knowledge about the canonical cell, such concepts are excluded from the GO. Finally, the GO includes almost 400 terms that are specific to particular species or classes of species, such as "viral capsid" (for virus) and "thylakoid" (for plants). Such terms are excluded from the FMA since they are not relevant for Homo sapiens. These two types of differences we assigned to the "domain specific" sub-type of breadth differences.

Just as for depth, it is interesting that neither ontology is broader than the other. The ontologies are again complementary, with each providing more breadth in some areas than the other.

The differences in the ontological organization category are not surprising. The FMA has more "is-a" ontology organization terms than the GO, while the GO includes more "part-of" organizational terms. We expected this result because the FMA is more structurally rigid and formalized than the GO [6]. Thus, the FMA uses organizational terms to maintain consistent use of the "is-a" relationships between terms, while the GO uses a mixture of "is-a" and "part-of" relationships to organize terms. Interestingly, the CCO also chooses to establish a more formalized "is-a" hierarchy than the GO even while borrowing many of its terms.

We recognize that our four categories of ontological differences may be somewhat ad hoc. However, we developed these over a significant period of iterative data analysis and with the help of biology and anatomy experts. In addition, our categories are consistent with the work of others who have compared large ontologies [7]. Our goal is to find a set of categories that can be used to analyze differences between any pair of ontologies.

Broader Implications and Conclusions

The process of creating inter-ontology mappings has several implications, particularly when it comes to querying across a domain area. Projects such as Biomediator [8] make use of an intermediate schema to translate queries across sources. However, this approach still relies on the ability of domain specialists to map their knowledge to this shared schema. Thus, understanding how to build better mappings across knowledge sources would have broad implications. Below, we discuss the implications for each of the four different sorts of differences that we found: language, depth, breadth, and organizational differences.

Language differences are by far the easiest and most straightforward to map. As these sorts of

terms represent a direct mapping between concepts, the only challenge is in determining synonymy between them. Mappings between synonymous terms across ontologies could lead to more complete synonym lists, which would allow for greater interoperability and ease of use by researchers.

It is more interesting to consider those terms which represent differences in depth of scope. Such terms typically map across ontologies via a common parent term. A proposed mapping would then consist of many-to-one mappings between the child terms and the common parent. Such mappings are inherently more complex than the one-to-one mappings for synonymous terms: They are non-symmetric, they must specify whether the child-parent relationship is "part-of" or "is-a", and they provide only related information, rather than truly linked information between ontologies.

When terms vary by breadth of scope, the challenge is even greater. In general, it will not be feasible to construct mappings for all differences in breadth of scope. For example, terms may be intentionally excluded by developers of either ontology as belonging outside the scope of that ontology. In our example, this is the case for species specific GO terms that should not be mapped to anything in the FMA. However, identification of these breadth differences may suggest areas for further refinement within an ontology. For example, the anatomists developing the FMA have expressed interest in including some GO terms that fall into this breadth difference category, such as protein complexes.

While terms which vary in scope may potentially provide some information across ontologies, terms referring to ontological organization are typically unique to a particular ontology. It seems unlikely that such organizational terms would be readily mapped across ontologies; they have no shared physical reality and they may require significant changes to the organization of many other terms.

There is established interest in creating mappings between biomedical ontologies, including work by Zhang and Bodenreider to automatically align the FMA with the GALEN Common Reference Model [7]. From our perspective, their work focuses on language differences. We find it interesting that in both comparisons (FMA vs GALEN or FMA vs GO) there are relatively few terms that can be connected by resolving these language differences: about 9% of the total number of terms. As we described above, differences in breadth and depth are much harder to align than language differences.

As collaborative research increases, and as the number of biomedical ontologies increases, we expect that there will be a growing need to analyze and understand the overlap and differences among ontologies. A use of our categorization of differences may be in the realm of semi-automatic mappings. If we can first categorize the sorts of differences, then it may be more clear which sorts are amenable to automatic methods rather than more manual work.

To summarize, by looking at cellular structure in the Foundational Model of Anatomy, the Gene Ontology, and the Cell Component Ontology, we arrived at four categories of differences. These categories have different implications for how easy it may be to establish linkages across the ontologies. We hope that our analysis will aid in the creation of more compatible ontologies, as well as assist with knowledge sharing and automated alignment between existing pairs of ontologies.

Acknowledgements

Thanks to Jose Mejino and Augusto Agoncillo for help with understanding the FMA and anatomy. This work was partially funded by the BISTI planning grant (#P20 LM007714) and a training grant (#T15-LM07442) from the National Library of Medicine.

References

1. Ogren PV, Cohen KB, Acquaah-Mensah GK, Eberlein J, Hunter L. The compositional structure of Gene Ontology terms. *Pacific Symp Biocomput*. 2005;174-185.

2. Rosse C, Mejino JL Jr. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. *J Biomed Inform*. 2003 Dec;36(6):478-500.

3. The Gene Ontology Consortium. Creating the gene ontology resource: design and implementation. *Genome Research*. 2001 Aug;11(8):1425-1433.

4. Zhang P, Paley S, Kaipa P, Rhee SY, Karp PD. MetaCyc and AraCyc. Metabolic Pathway Databases for Plant Research. *Plant Physiology*. 2005;138:27-37.

5. Gennari JH, Musen MA, Fergerson RW, et al. The evolution of Protégé: an environment for knowledge-based systems development. *Intl J of Human-Comp Studies*. 2003 Jan;58(1):89-123.

6. Smith B, Rosse C. The role of foundational relations in the alignment of biomedical ontologies. *Medinfo*. 2004 Sep;444-448.

7. Zhang S, Bodenreider O. Aligning representations of anatomy using lexical and structural methods. *Proceed of AMIA Symp.* 2003;753-757.

8. Shaker R, Mork P, Barclay M, Tarczy-Hornoch P. A rule driven bi-directional translation system for remapping queries and result sets between a mediated schema and heterogeneous data sources. *Proceed of AMIA Symp*. 2002;692-696.