Definition Coverage in the OBO Foundry Ontologies: The Big Picture

Daniel R. Schlegel
and Peter L. Elkin
Department of Biomedical Informatics
University at Buffalo, SUNY, Buffalo, NY, USA
Email: <drschleg, elkinp>@buffalo.edu

Selja Seppälä
Department of Health Outcomes and Policy
University of Florida, FL, USA
Email: sseppala@ufl.edu

I. Introduction

High quality ontologies have both textual and logical definitions for their terms. Definitions serve many purposes: good textual definitions allow for experts and non-experts alike to understand the content of an ontology and use it in the manner the authors intended; logical definitions are necessary for reasoners to verify that an ontology is consistent, and may make application of the ontology easier for users. Ideally, logical and textual definitions would convey the same information, and each can provide an accuracy check on the other [1], [2].

Producing definitions is difficult and time-consuming. Thus, despite the best efforts of ontology developers and the existence of a number of tools and methods to populate ontologies with definitions, it is not uncommon to see missing textual or logical definitions, if not both. This is also the case in the Open Biomedical Ontologies (OBO) Foundry [3] ontologies.

The OBO Foundry contains 9 'core' ontologies and 128 non-core ontologies.[1] These ontologies are developed in a coordinated way according to a set of shared principles.[2] One of the OBO Foundry principles is about definitions: member ontologies should have "textual definitions ... for a substantial and representative fraction [of terms], plus equivalent formal definitions (for at least a substantial number of terms)."[3] The statement of this principle is rather vague and elicits an obvious question: How much is 'substantial'?

We examine the coverage of textual and logical definitions throughout the OBO Foundry ontologies. In particular, we aim to determine: (1) if the prevalence of definitions is different between the core and non-core ontologies; (2) if there are more textual than logical definitions; (3) if the size of ontologies has an effect on definitional coverage. To conclude, we discuss ways of quantifying the notion of 'substantial' definition coverage to determine to what extent the principle of having textual and logical definitions for a substantial number of terms is upheld.

[1]Our study focuses on 119 ontologies out of the 137 present in the OBO Foundry, since 18 non-core ontologies were either unavailable on the web due to broken links, or they failed to load using the OWL API.

[2]http://obofoundry.org/principles/fp-000-summary.html.

[3]http://obofoundry.org/principles/fp-006-textual-definitions.html.

II. Methods

Textual definitions tell us about the properties of the instances of a class in an ontology. They typically have two parts: (i) a *genus* that states the type of thing of which they are instances, and (ii) one or more *differentia(e)* that state the properties of these instances that differentiate them from instances of neighboring types.

To identify textual definitions, we used the IAO annotation property `definition` used in 103 of the 119 ontologies in this study. We also examined the set of annotation properties used in the OBO Foundry ontologies that contained the string `def` but did not contain the strings `editor`, `source`, `citation`, `defines`, or `defined` to try to capture any non-standard annotation properties which might have been used to signal a definition. We also included the IAO annotation property `elucidation` for ontologies that contain some primitive classes that cannot be, strictly speaking, defined.

One of the main components of ontologies are classes, which are defined by *class expressions*. Class expressions represent conditions that individuals must satisfy to be members of a class. Some axioms, such as `SubClassOf` and `EquivalentClass`, define relationships between class expressions. These two axiom types, specifically, constitute the logical definitions of the ontology terms. We say an axiom contains a genus for the definition of class c_1 if the axiom contains some other class, c_2, where c_2 is not part of an object property restriction; an axiom contains one or more differentiae for the definition of a class if the axiom contains any object property restrictions.

For each ontology, we computed the number of classes that contain: (i) at least one genus; (ii) at least one differentia; and (iii) at least one of both. A class specified by both a genus and one or more differentiae has a *complete logical definition*.

III. Results

We review our results in light of our goals stated in section I.

Item (1): Table I shows that the prevalence of definitions is different between the core and non-core ontologies. We found that coverage within the 9 core ontologies was quite high, with 6 having textual definitions for over 90% of their terms. On average, core ontologies have textual definitions for 85.6% of their terms ($stdev = 21\%$); non-core ontologies, 63%

TABLE I

COVERAGE OF TEXTUAL DEFINITIONS, LOGICAL DEFINITIONS, AND PARTS OF LOGICAL DEFINITIONS ACROSS THE CORE, NON-CORE, AND SUM TOTAL OF THE ANALYZED ONTOLOGIES IN THE OBO FOUNDRY.

	Core	Non-Core	Total
Textual Definition Coverage	86%	64%	66%
Logical Definition Coverage	53%	28%	30%
Genera Covereage	91%	86%	86%
Genera Only Coverage	39%	58%	57%
Differentiae Covereage	53%	34%	36%
Differentiae Only Covereage	0%	6%	6%

($stdev = 38\%$). Coverage for complete logical definitions among the core ontologies was 53% ($stdev = 34\%$), and only 28% ($stdev = 29\%$) for the non-core ontologies. Over the full set of analyzed OBO Foundry ontologies, textual definition coverage is on average 66% ($stdev = 37\%$) and complete logical definition coverage, 30% ($stdev = 30\%$).

Item (2): Figure 1 shows that the studied ontologies have more textual than logical definitions and that the trends are nearly opposite. Relatively few ontologies have poor textual definition coverage, while a large number have 90-100% coverage. Conversely, a large number of ontologies have very poor logical definition coverage (0-10%), and few have good logical definition coverage.

Item (3): Figure 2 shows a correlation between ontology size and logical definition coverage. We grouped the ontologies as follows: 'very small' (0-99 terms, n=17); 'small' (100-999, n=42); 'medium' (1,000-9,999, n=44); 'large' (10,000-99,999, n=11); and 'very large' (100,000+, n=3). We found that nearly all groups had textual definitions for roughly 60-70% of their terms. The 'large' category formed the only outlier, with a 33% coverage. We examined logical definition coverage in three ways — the percent of classes: with genera; with differentiae; and with both. The percent coverage of complete logical definitions rose slowly as ontology size grew.

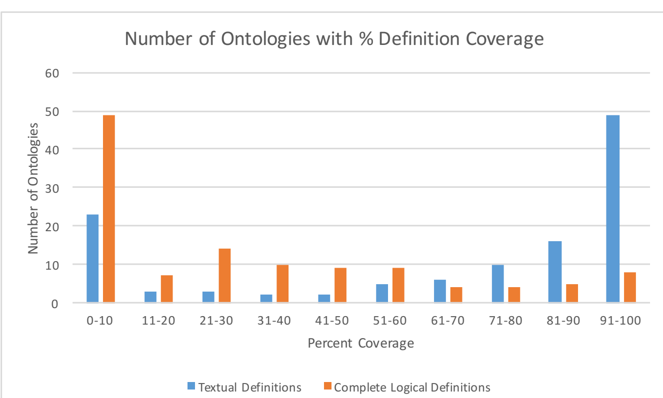

Fig. 1. The number of ontologies with percent coverage of textual and complete logical definitions.

IV. Discussion and Conclusion

Determining if the principle of having textual and logical definitions for a substantial number of terms is upheld requires quantifying the notion of 'substantial' definition coverage. If we consider that 'substantial' equates with the average definition coverage measured over the core ontologies, then an adequate coverage to be included in the OBO Foundry would be to have at least 86% of the terms specified with a textual definition and 53% with a complete logical definition. Whereas, considering all of the (analyzed) ontologies in the OBO Foundry, we get, respectively, 66% and 30%.

To expect that all ontologies have coverage as complete as the core ontologies is unrealistic. Therefore, we quantify 'substantial' at roughly 65% for textual definitions, and propose that logical definitions be held to this standard as well.

Having set a measure for substantial definition coverage in the OBO Foundry ontologies, our results show that on average there is substantial coverage of textual definitions, but not of logical definitions.

Definitions, both logical and textual, are essential components of an ontology. The OBO Foundry has the noble goal of creating a repository for ontologies developed using a shared set of principles, including some (vague) requirements for including definitions. This study is the first one not only to analyze the "big picture" of definition coverage in the OBO Foundry, but also to suggest a numeric value for 'substantial' definition coverage.

References

[1] S. Seppälä, Y. Schreiber, and A. Ruttenberg, "Textual and logical definitions in ontologies," in *Proceedings of DIKR 2014, IWOOD 2014, and OBIB 2014*, Boyce, R., et al., Ed., vol. Vol-1309. Houston, TX, USA: CEUR Workshop Proceedings (CEUR-WS.org), October 6-7 2014, pp. 35–41. [Online]. Available: http://ceur-ws.org/Vol-1309/

[2] S. Seppälä, Y. Schreiber, A. Ruttenberg, and B. Smith, "Definitions in ontologies," *Cahiers de lexicologie*, vol. 4, no. Numéro thématique "Au coeur de la définition", forthcoming.

[3] B. Smith, M. Ashburner, C. Rosse *et al.*, "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration," *Nature biotechnology*, vol. 25, no. 11, pp. 1251–1255, 2007.

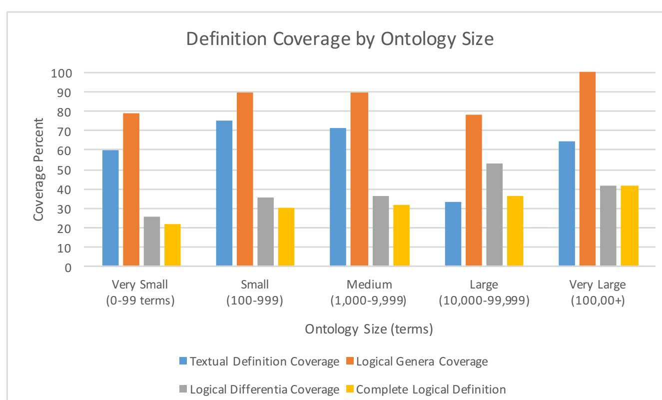

Fig. 2. The coverage of textual and logical definitions by ontology size. Both the genus and differentia components of the logical definitions are shown, along with coverage for the complete logical definitions.