An epistemology for an ontology of biomedical relations

Cohen, (unordered) Verspoor, Zweigenbaum, Grimpe, Yadav, Bada, (senior) Hunter
Affiliation: dept. name of organization, name of organization, acronyms acceptable, City, Country

Abstract—**The Relation Ontology has not had the expected impact on biomedical research. Prior results were consistent with the hypothesis that this is due to a coverage issue: the Relation Ontology is not used to its full potential because it lacks many relations that would be necessary to represent the reality of the biomedical domain. This paper presents a rationalized approach to increasing the coverage of the Relation Ontology by analysis of terms in the existing Open Biomedical Ontologies.**

Keywords: ontology; ontologies; relations; Relation Ontology; epistemology of ontology

I. Introduction

Ontologies of the biomedical domain have been one of the greatest achievements of the genomic era. They have enabled enormous advances in a wide variety of fields of research, e.g. model organism database curation and high-throughput assay analysis.

Ontologies are based on relations between concepts. In 2005, Smith et al. made a crucial advance in the field of biomedical ontologies by presenting an ontology of the very relations around which ontologies themselves are constructed [3]. Its evolution is interesting in that it started out as a very small and constrained theory of relations that tried to to model a set of primitives; however, over the years since its original publication, it has become quite large. Rather than aiming to be constrained and/or to propose a set of primitives, it now proposes a large and nuanced set of relations. It retains the goal of being a realistic model of “what is” in biomedicine, but in the absence of its original constraint (in the sense above), its epistemology is now unclear.

Despite the clear importance of the purposes of the Relation Ontology, its impact, as measured by its use in other ontologies and by its count of citations, has been small. This is a surprise. A previous pilot study that attempted to understand why its impact has been limited found results that are consistent with the notion that its impact has been limited because its *coverage* is low: despite its large size, it mostly lacks the relations that would be necessary to model the relations that characterize the ontology (in its sense of a field of study) of the biomedical domain[1].

The current situation can be summarized as follows: ontologies are important to modern biomedical science; the Relation Ontology is important to ontologies; but, the low coverage of the Relation Ontology has interfered with it making as many contributions as it could. This paper therefore evaluates a methodology for increasing that coverage, and for doing so with a defined epistemology, in the sense of a specification of what would count as evidence for a proposed relation.

The methodology that is proposed here for increasing the coverage of the Relation Ontology is to use the terms from biomedical ontologies to drive a proposed set of additions to the Relation Ontology. Briefly, the approach is to find substrings of terms in ontologies, and then to propose those to the Relation Ontology maintainers as potential relations.

A potential argument against the proposed method might go as follows. The claim would be that the goal of an ontology is to model *what is,* and that this methodology is modeling how people *talk about* what is.

This objection would be inaccurate. The proposal is not to model the language that people use to talk about what is; neither does it propose to build ontologies from scientific texts. Rather, the proposal is to use language as a form of *evidence* for “what is”--in essence, to provide epistemological support for extension of the Relation Ontology. Using the Open Biomedical Ontologies as the input data has the advantage of remaining focused on the biomedical domain. There is no claim made or implied here that this is a

[1] An ontologist who is quite familiar with the Relation Ontology mapped two other sets of relations to the Relation Ontology. One set of relations draws its motivation and methods of construction from linguistics, and is quite constrained--a model of semantic primitives. The other set of relations is very domain-specific, with the goal of being a necessary but not sufficient set of relations for doing information extraction from biomedical literature. Two thirds of the small set of theoretically-motivated relations could be mapped to the Relation Ontology; none of the larger set of domain-specific relations could be mapped to the Relation Ontology, despite the fact that it is ten times the size of that small set of relations, and that those additions to the original primitives are apparently intended to be domain-specific: a decade of development of this ontology has left it unsuited for any application that we would be interested in.

methodology that would yield a complete set of relations--indeed, there is reason to think that we will *never* have a complete set of relations [3]. The claim is not that the methodology would, or could, result in a *sufficient* set of relations--only that it would result in a *necessary* set of relations.

5. Materials and Methods

A. Materials

Relation Ontology version of 2017-04-11.

Gene Ontology version 2017-09-12.

B. Methods

All 47,047 terms were filtered to remove obsolete terms and terms containing brackets or parentheses, leaving 43,360 terms in total. Barnbrook's algorithm [1] was applied from the left edge to a depth of one token. The resulting strings were ordered by frequency. A linguist reviewed them to remove adjectives and to convert deverbal nouns to verbs. Two biologists (BG/PY) reviewed the results for biological validity, a terminology expert (PZ) screened them for terminological validity, and two ontologists (LEH/MB) reviewed them for ontological validity. We then attempted to map them to the Relation Ontology.

All code, the version of the Gene Ontology that was used for the experiment, and all output files are available on GitHub at https://goo.gl/zKtN4J.

II. Results

The method yielded 9,914 unique strings, of which 3,594 appeared more than once. For space, only the top ones are listed here. We give the count for each, and indicate whether or not it was mappable to the Relation Ontology.

TABLE I. Top outputs for the Gene Ontology

Barnbrook algorithm output	*Count*	*In RO?*
regulate	3311	Yes
respond	570	-
modulate	272	-
detect	172	-
establish	148	-
maintain	109	-
suppress	103	-
signal	78	-

Table 1 shows that only the single most frequent of those strings could be mapped to the Relation Ontology. All others are biologically valid, but are absent from the Relation Ontology.

III. Discussion and Conclusions

The Relation Ontology started out as a very constrained theory of relations that tried to to model a set of relational primitives. In this, it was very similar in goals and in its resulting content to what a semanticist do to represent the semantics of verbs. But, that is not what is being done here--the goal of this work is to use ontologies to propose relations that are part of *what is,* not of verbs (or of any other aspect of language).

The facts that the method found *regulate* as its top hit and that the Relation Ontology includes *regulate* is consistent with the proposal that this technique can find relations that belong in the Relation Ontology. (It is notable that this relation was added to the Relation Ontology after first being added to the Gene Ontology, and that it was added to the Gene Ontology as a result of a linguistic analysis of Gene Ontology terms.) The fact that all but one of the top hits are not currently represented in the Relation Ontology is consistent with the idea that this methodology could yield a significant increase in Relation Ontology coverage.

From the perspective of the Relation Ontology and its goal of modeling not language, but rather what is in the domain, we note that the set of relations that are suggested by the technique that is used here does not propose *anything* about language. Rather, it suggests only relations that are about the domain. There are many other techniques that could be applied to terms from Open Biomedical Ontologies--future work should include exploring them.

Author contributions: KBC conceived of the idea, did the initial analysis, and wrote the first draft of the paper. MB ensured ontological validity. BG and PY ensured biological validity. PZ ensured terminological validity. KV and PZ ensured epistemological validity. LEH guided the work. All authors contributed to or reviewed the analysis and participating in the writing or reviewing of the paper.

Acknowledgments

Funding agencies.

References

1. Barnbrook, Geoff. Defining Language: A local grammar of definition sentences. Vol. 11. John Benjamins Publishing, 2002.
2. Nastase, Vivi, et al. "Semantic relations between nominals." Synthesis lectures on human language technologies 6.1 (2013)
3. Smith, Barry, et al. "Relations in biomedical ontologies." Genome Biology 6.5 (2005): R4