Where GO is going and what it means for ontology extension

Catia Pesquita and Francisco M. Couto

Departamento de Informática, Faculdade de Ciências da Universidade de Lisboa
Campo Grande, Lisboa, Portugal

Abstract. Developing and maintaining a biomedical ontology is a time and effort-consuming task, given the dynamic and expanding nature of biomedical knowledge. This is a relevant issue for very large ontologies which cover a broad domain, for smaller ontologies maintained by a small team and also for domains where being able to perform quick updates is critical (e.g. epidemiology).

The first step in the process of extending an ontology is identifying the areas of the ontology that need to be changed - change capturing. In this paper we propose that this process can be semi-automated by exploring ontology information. This would be a valuable tool to support ontology developers in ontology extension, easing their burden.

In order to accomplish this, we have developed a framework for analysing the extension of ontologies, to create a general panorama of ontology extension processes that can guide the development of change capturing techniques. We have applied it to the analysis of the extension of the Gene Ontology and uncovered some of the underlying tendencies in its extension. Building upon the results of this analysis and a set of guidelines for ontology change capturing, we then investigated the feasibility of prediciting which classes of the ontology will be extended in a future version.

Finally, we discuss the obtained results and indentify the main challenges and future directions for the budding area of ontology extension prediction.

Key words: Biomedical ontologies, ontology extension, ontology refinement, ontology enrichment, prediction of ontology extension

1 Introduction

The development of a biomedical ontology is a very demanding process that requires both expertise in the domain to model and in ontology design. It is also necessarily an iterative process [10] since biomedical knowledge is diverse, complex and continuously changing and growing. This process, usually named ontology evolution, requires large investments of both time and money with each new ontology version that is produced.

Ontology evolution can be defined as the process of modifying an ontology in response to a certain change in the domain or its conceptualization [6]. These include changes in the portion of the real world they model, the uncovering of information previously unavailable, a reassessment of the relevance of some element to the ontology or a need to correct previous mistakes [4]. In the last couple of years, a generally agreed upon model for ontology evolution has emerged, which includes four distinct steps: (1) requesting the change, (2) planning the change, (3) implementing the change and (4) verification and validation (for a review see [8]). The changes made in the course of ontology evolution can be of three elementary types: addition, removal and modification [13]. We define ontology extension as the process of ontology evolution concerned with the addition of new elements. We consider ontology extension to encompass both ontology refinement (the addition of new classes to an ontology) and ontology enrichment (the addition of non-taxonomical relations or richer axioms).

Before these changes are actually performed, the need for the change must be identified. This is the first step in ontology evolution, the change capturing phase [12], and it can be based on explicit or implicit requirements [5]. Explicit requirements correspond to those made by the ontology developers or to requests made by end-users. Implicit requirements correspond to those that can be uncovered by change discovery. Stojanovic et al. [13] list a series of guidelines for change capturing, organized into three types according to the kind of data they exploit, to which Castaño et al.[2] add a fourth:

structure-driven: which are derived from the structure of the ontology, e.g. ‘A class with a single subclass should be merged with its subclass’.

data-driven: which correspond to implicit changes in the domain and are discovered through the analysis of the instances belonging to the ontology, e.g. ‘A class with many instances is a candidate for being split into subclasses and its instances distributed among newly generated classes’.

usage-driven: which are deduced from the usage patterns of the ontology in the knowledge management system e.g. classes that have not been retrieved in a long time might be out of date.

discovery-driven: which is applied when a new instance cannot be described by the ontology classes, and new classes are identified using external resources.

These changes can in principle be semi-automatically discovered by analyzing the ontology data and its usage. This process can be cast in terms of a prediction of ontology extension and can represent a significant contribution to easing the burden of keeping an ontology up-to-date.

The main application of a methodology to predict ontology extension is to support manual or semi-automated ontology extension, since it can minimize the effort in collecting and crossing the information necessary to make the extension decisions. It can contribute to the evolution of larger ontologies by helping to pinpoint the areas that are in need of attention and also to smaller ontologies, where team size and resources may not be as large, by reducing the time and effort spent. It can also provide valuable assistance when there is an urgent need to extend an ontology to cover a new aspect, such as in the case of an epidemics, where the inclusion of new classes in a timely fashion can improve the performance of data analysis methods [11]. Furthermore, they can be incorporated into automated and semi-automated ontology extension systems, which so far have not addressed this issue [7][9][14].

To guide the development of methods for automated prediciton of ontology extension it is of interest to analyze the extension of ontologies. In [4], a methodology for calculating the improvements obtained in successive versions of biomedical ontologies based on the matches and mismatches between them is proposed. It has been used to calculate the degree of correctness of the Gene Ontology (GO) terminology and to forecast how this overall quality will improve [3]. In this detailed analysis of the evolution of GO 75% of the changes made to classes were identified as being insertions. However this study has a strong focus on error correction, which includes not only the addition of new elements, but also their removal: of the 17 parameters used, only two correspond to the absence of elements that should be included in the ontology. Furthermore, the method does not take into account the hierarchical level at which the error is made, nor does it consider GO annotations.

In this paper we present a preliminary framework for analysing the extension of an ontology, and apply it to the analysis of GO. We have chosen GO for our study because it presents a very interesting case: it is the most prominent bio-ontology, with widespread use and impact; it covers a considerable wide domain; it provides a corpus of annotations and it is updated on a frequent basis, thus supporting the investigation of its evolution through the analysis of different versions. Building upon the analysis of ontology extension, we then investigate the feasibility of predicting the evolution of GO. Since GO authors do not justify the addition of new elements, we based our predicition in a set of rules derived from the guidelines for ontology change capture. We were interested in evaluating the suitability of these guidelines as a support for ontology extension prediction. Finally, we discuss the novel area of ontology extension prediction, its challenges and role in the future of ontology development.

2 A Framework for Analyzing Ontology Extension

An analysis of ontology extension should by definition, focus on both refinement and enrichment, and analyze several versions of the same ontology during a time period. The decision on the time period to analyze should be based on the age of the ontology as well as the availability and frequency of new version releases.

For analyzing ontology refinement we propose inspecting three key aspects:

1. depth of new classes, i.e. minimum distance to the root class over *is_a* and *part_of* relations.
2. number of new classes that are children of existing *vs.* newly added classes
3. number of new classes that are children of leaf classes

The first and third aspects capture the general direction of the refinement of the ontology, where additions at a greater depth and to leaf classes represent vertical extension whereas additions at middle depth and to non-leaf classes represent horizontal extension. These aspects are helpful to analyze the level of detail and coverage provided by the refinement. The second aspect is related to another interesting characteristic of refinement, whether new classes are inserted individually or whether as part of a new branch.

For analyzing ontology enrichment we propose investigating the following:

4. age and depth of the classes linked by the new relation (i.e. whether the relation is established between old classes, between an old and a new class or between new classes)

This aspect is intended to capture first at what level of specificity do the enrichment events take place, and secondly if enrichment happens alongside refinement or if it succeeds it.

2.1 Analyzing the Gene Ontology Extension

Based on the aspects identified in the previous section, we have analyzed 12 versions of the Gene Ontology and its annotations equally spaced over a period of 6 years (2005-2010). At 6 month intervals, new classes represent about 5% of all classes in the ontology. In the context of GO, enrichment corresponds to the insertion of new non *is_a* relations between existing or newly inserted classes.

Figures 1, 2, 3 and 4 show the results of the analysis of each aspect. In all three hierarchies, the majority of new subclasses are added as children of non-leaf classes, resulting in a prevalence of horizontal extension. Also, the refinement of molecular function and cellular component occurs mostly via single insertions, whereas in the biological process groups of related classes are inserted together. Regarding enrichment, in biological process, a considerable portion of relations are established between two newly inserted classes, whereas in cellular component, the majority is made between an existing and a new class.

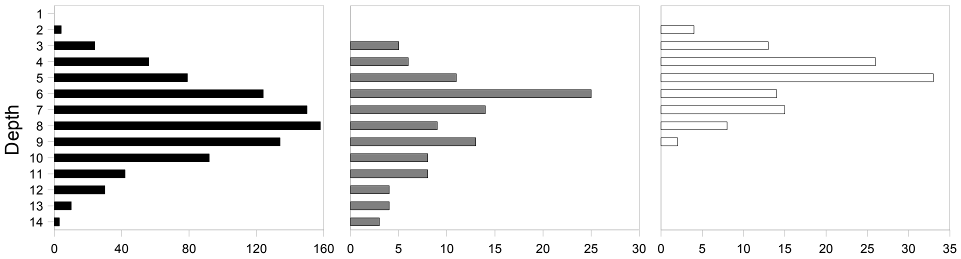

Fig. 1. Average depth of new classes

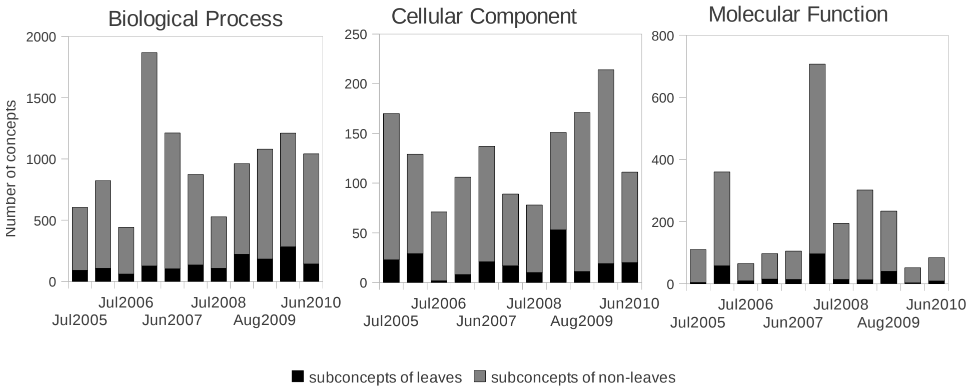

Fig. 2. Ancestry of new classes (leaves or non-leaves) by ontology version

3 Predicting Ontology Extension: a rule-based approach

Adapting and extending the guidelines proposed by [13] following [10] that are concerned with ontology extension, we recognize two heuristics to identify potential ontology extensions and classify them according to the type of data they use:

1. structure-driven: If a class has fewer children than its siblings, it may be a candidate for extension
2. data-driven: A class with many instances is a candidate for being split into subclasses and its instances distributed among newly generated classes.

Following the above mentioned guidelines we have devised a set of rules to apply to the prediction of the extension of GO. The rules aim at finding a partition of the set of classes that best separates classes that will be refined in a future version from those that will not. Here, we assume that the latest ontology version is believed to be as correct and complete as possible [4]. We have three types of rules, one structure-based and two data-based. The structure-based rules are derived from guideline 1:

Rule 1: A class with at most less $x\%$ subclasses than its siblings is a candidate for refinement

with x taking four evenly spaced values between 25 and 100%. The data-based rules are derived from guideline 2 but distinguish between the set of all annotations and the set of manually curated ones:

Rule 2: A class with at least more $x\%$ annotations than its siblings is a candidate for refinement

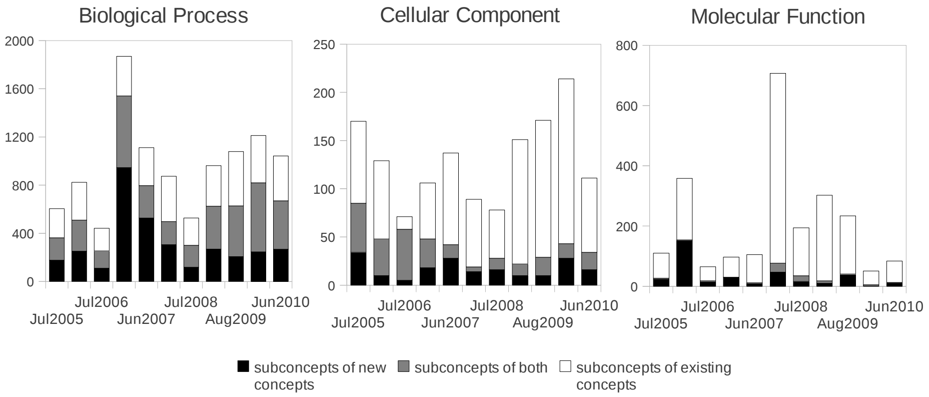

Fig. 3. Ancestry of new classes (existing or new parents) by ontology version

Rule 3: A class with at least more $x\%$ manual annotations than its siblings is a candidate for refinement

with x taking four evenly spaced values between 100 and 250%.

Distinguishing between these two sets of annotations is very relevant in the context of GO, since the set of manual annotations contains only those that have been reviewed by a curator and can therefore be considered more reliable. Nevertheless, only about 3% of all annotations are manual which means they provide a narrower coverage.

We applied these rules to classes across the 12 ontology versions. To accomplish this, we checked how well the two sets of classes created by the application of each rule, reflected the sets of classes that were refined and not refined in a future version at 6 months, 1 and 2 years. To evaluate the predictive power of the rules, we computed the number of true positives, true negatives, false positives and false negatives, and used the following indicators:

$$precision = \frac{tp}{tp+fp} \qquad recall = \frac{tp}{tp+fn}$$

$$f - measure = 2 \times \frac{precision \times recall}{precision + recall}$$

Table 1 shows these results for refinement in 6 months [1] for the values of x that generated the best results.

Although these results are overall poor, there is a marked difference between the performance of structure and data-based rules, with data-based rules having a higher precision for all 3 hierarchies and a higher recall in molecular function.

[1] results for 1 and 2 years were similar - data not shown

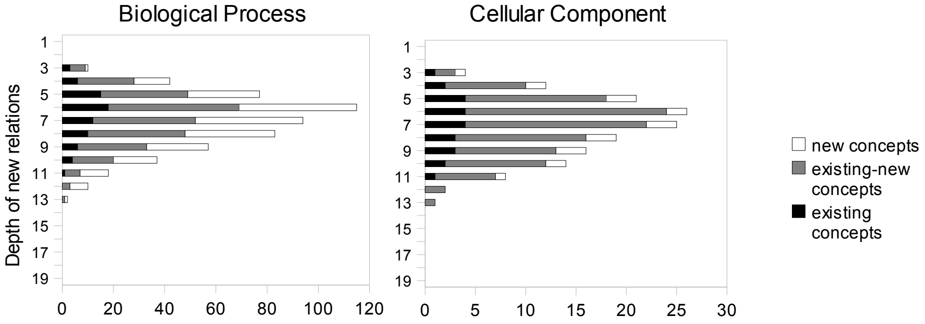

Fig. 4. Depth and age of the classes in new enrichment relations. Molecular Function not shown, since it contains less than 10 non *is_a* relations.

We also applied these rules to prediciting the refinement for ontology branches as a whole, as opposed to the previous strategy that predicted refinement for individual classes. This follows from the observation that many of the new classes inserted in the biological process hierarchy are inserted as part of small subgraphs rather than single insertions. We focused on the subgraphs that are rooted on classes at a depth of 4 due to the fact that most extension events occur at this depth or lower. However, the results obtained were comparable to those generated by predicting for individual classes.

4 Discussion

The application of our framework for ontology extension analysis to GO has yielded some interesting results. Firstly, the majority of new classes are not added to leaf classes, resulting in a horizontal growth of the ontology. This means that GO is not adding increasingly specific classes but rather fleshing out. Secondly, we have identified that in GO refinement happens by two major modes: individual insertions and group insertions. The first occurs frequently in all GO hierarchies, whereas the second is only common in the biological process hierarchy. This is in line with the fact that most of GO's special interest groups belong to the biological process area and their work is more focused on modelling portions of their areas of interest rather than making individual insertions. We are aware that our usage of path-based depth to define the sub-graphs of GO that are subject to extension, can suffer from bias, since terms at the same depth do not necessarily express the same degree of specificity [1]. However, we have decided to use path-based depth, since we needed to create sub-graphs independently of their number of annotations, so as not to introduce a bias to our annotation based rules.

Biological Process

Rule	Precision	Recall	F-measure
1 ($x = 75\%$)	0.0772 ± 0.0317	0.364±0.0802	0.127±0.0479
2 ($x = 200\%$)	0.220±0.0185	0.318±0.0638	0.256±0.0128
3 ($x = 200\%$)	0.242±0.0302	0.380±0.0507	0.292±0.01714

Cellular Component

Rule	Precision	Recall	F-measure
1 ($x = 75\%$)	0.0270±0.0228	0.381±0.206	0.0501± 0.0406
2 ($x = 200\%$)	0.119±0.109	0.212±0.246	0.149 ±0.148
3 ($x = 200\%$)	0.199 ±0.121	0.374±0.259	0.252 ± 0.156

Molecular Function

Rule	Precision	Recall	F-measure
1 ($x = 75\%$)	0.0122±0.0033	0.223±0.0908	0.0230±0.0060
2 ($x = 200\%$)	0.101±0.0388	0.406 ±0.0357	0.157 ±0.0492
3 ($x = 200\%$)	0.123 ±0.0515	0.526±0.0573	0.194±0.0672

Table 1. Prediction results for the refinement of the Gene Ontology at 6 months. Shown values are averaged over all ontology versions, resulting from a total of 11 runs.

This refinement by branches in the biological process hierarchy is also captured by the enrichment analysis, where there is a high proportion of new enrichment relations that are established between new classes.

Theoretically, these two modes of refinement should impact semi-automated change capturing methods, hence we applied the rules for both individual and branch extension predicition. However such impact was not visible, likely due to the poor performance obtained.

These results emphasize that the current proposed guidelines for capturing change based on structure and data are not appropriate for handling a large and complex ontology such as the Gene Ontology. We are aware that the guidelines represent an effort to ensure a balanced structure for the ontology, and that given the size and evolving nature of the domain GO covers, its extension cannot be governed alone by these precepts. In fact, GO's Ontology Development group [2] has highlighted the processes used in the identification of areas that need to be developed:

- by working closely with the reference genome annotation group to ensure that areas that are known to undergo intense annotation in the near future are updated
- by listening to the biological community
- by ensuring that emerging genomes have the necessary classes to support their needs

If GO's change management regarding extension were to be made explicit, for instance as is the case for making a term obsolete where the reason is given, we

[2] http://wiki.geneontology.org/index.php/Ontology_Development_group_summary

could perform a more in-depth analysis and perhaps derive more accurate rules. Nevertheless we have obtained better results using the number of annotations rather than the number of subclasses, which may be related to the fact that GO development is driven by need, which can be approximated by the rate of annotation, rather than by a process of homogeneization of structure. In fact, this difference was to be expected considering that in GO's domain the level of specificity of each branch is dependent on natural and scientific phenonmena, which prevents the existance of an homogenous structure to the ontology. Such structure-based guidelines are however expected to function better in ontologies that follow a more conceptual approach.

In trying to predict ontology extension, particularly in the case of large biomedical ontologies, we are facing a multitude of variables, not only the advancement of biomedical knowledge and the current state of the ontology itself but also social and technical aspects. The extension of biomedical ontologies occurs via several different processes, and motivated by distinct needs, which cannot be apprehended by a 'one size fits all' rule. We believe that to handle this complexity, we need to employ more sohpisticated techniques, that are able to handle numerous variables and more complex relations between them.

Therefore we are currently working on a supervised learning methodology to support the prediciton of ontology extension that explicitly addresses these issues.

We believe that the future of ontology development will necessarily incorporate the automation of some of its processes, mainly those that are tedious and time-consuming, releasing ontology experts to focus on core modelling issues. We have outlined one of these processes, semi-automated change capturing via prediction of ontology extension and presented some of the issues and challenges in this budding field.

5 Acknowledgements

This work was supported by FCT, through the Multiannual Funding Programme and the PhD grant SFRH/BD/42481/2007. The authors also wish to thank the European Commission for the financial support of the EPIWORK project under the Seventh Framework Programme (Grant #231807)

References

1. G. Alterovitz, M. Xiang, D. P. Hill, J. Lomax, J. Liu, M. Cherkassky, J. Dreyfuss, C. Mungall, M. A. Harris, M. E. Dolan, J. A. Blake, and M. F. Ramoni. Ontology engineering. *Nature biotechnology*, 28(2):2008–2011, 2010.
2. S. Castano, A. Ferrara, and G. Hess. Discovery-driven ontology evolution. *The Semantic Web Applications and*, 2006.
3. W. Ceusters. Applying evolutionary terminology auditing to the Gene Ontology. *Journal of biomedical informatics*, 42(3):518–29, 2009.

4. W. Ceusters and B. Smith. A realism-based approach to the evolution of biomedical ontologies. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, (4):121–5, Jan. 2006.
5. P. Cimiano and J. Völker. A framework for ontology learning and data-driven change discovery. *Proc. of the NLDB2005*, 2005.
6. G. Flouris, D. Manakanatas, H. Kondylakis, D. Plexousakis, and G. Antoniou. Ontology change: classification and survey. *The Knowledge Engineering Review*, 23(02):117–152, 2008.
7. J.-B. Lee, J.-j. Kim, and J. C. Park. Automatic extension of Gene Ontology with flexible identification of candidate terms. *Bioinformatics (Oxford, England)*, 22(6):665–70, 2006.
8. P. D. Leenheer and T. Mens. Ontology evolution: State of the Art and Future Directions. 2(1):1–47, 2008.
9. V. Novácek, L. Laera, S. Handschuh, and B. Davis. Infrastructure for dynamic knowledge integration–automated biomedical ontology extension using textual resources. *Journal of biomedical informatics*, 41(5):816–28, Oct. 2008.
10. N. F. Noy and D. L. Mcguinness. Ontology Development 101 : A Guide to Creating Your First Ontology. *Development*, pages 1–25, 2000.
11. F. Silva, M. Silva, and F. Couto. Epidemic Marketplace: an e-Science Platform for Epidemic Modelling and Analysis. *ERCIM News*, 2010.
12. L. Stojanovic, A. Maedche, B. Motik, and N. Stojanovic. User-driven Ontology Evolution Management. *Proceedings of the 13 th International Conference on Knowledge Engineering and Knowledge Management (EKAW-02), Lecture Notes in Computer Science (LNCS), Volume 2473, Springer-Verlag*, pp:285–300, 2002.
13. L. Stojanovic and B. Motik. Ontology Evolution Within Ontology Editors. *Proceedings of the OntoWeb-SIG3 Workshop*, pp:53–62, 2002.
14. T. Wächter and M. Schroeder. Semi-automated ontology generation within OBO-Edit. *Bioinformatics (Oxford, England)*, 26(12):i88–96, June 2010.