Quality evaluation framework for bio-ontologies

Jesualdo Tomas Fernandez-Breis[1], Mikel Egaña Aranguren[2], Robert Stevens[2]
[1]University of Murcia, Murcia, Spain; [2]University of Manchester, Manchester, UK

Abstract

Over the past few years the number of bio-ontologies has rapidly increased. The evaluation of ontologies has long been a problematic issue. The growing number of ontologies makes the need for a strategy for evaluating quality more urgent. We propose a framework for evaluating the quality of bio-ontologies. This framework is inspired by a well-known software quality standard, which has been adapted to the needs of ontology evaluation. An example of how to use the framework, comparing two versions of the Open biomedical Ontologies' Cell Type Ontology, is included as an illustration.

Introduction

Bio-ontologies have increased in number and importance since the development of the Gene Ontology. Many research groups are collaborating in the development of an orthogonal collection of bio-ontologies, the Open Biomedical Ontologies (OBO) Foundry (http://www.obofoundry.org). In addition, there also exist independent efforts for developing other bio-ontologies. The development of application ontologies, for example, usually requires the reuse of different ontologies, so bits from different ontologies have to be combined. For this purpose, developers have to decide which ontology to use, but they lack support for making an informed decision. Hence, there is a clear need for methods for evaluating the quality of bio-ontologies. Ontology quality evaluation has usually been the concern of the Ontology Engineering community, and has been addressed from different perspectives and hence related work in ontology evaluation can be classified according to the particular evaluation aim: ranking, correctness, or quality.

Ontology Engineering has historically adapted methods from the Software Engineering field since they have many stages in common. Recent examples are ontology development methodologies[1] or Ontology Design Patterns[2]. There has not, however, been any attempt to adapt Software Engineering approaches for evaluating ontology quality. In this work, we propose an evaluation framework for bio-ontologies that is inspired by the ISO 9126 (http://en.wikipedia.org/wiki/ISO_9126) standard for software quality, which has been applied in other fields, for different purposes, such as the evaluation of e-learning systems[3] or software design documents[4]. Its application is recommended because: (1) it provides a comprehensive specification and evaluation model for software product quality; (2) it addresses user needs of a product by allowing for a common language for specifying user requirements that is understandable by users, developers and evaluators; (3) it objectively evaluates quality of software products based on observation; and (4) it makes quality evaluation reproducible. All these properties are desirable for an ontology quality evaluation approach, and hence they represent a potentially useful tool e such a framework.

Furthermore, this standard does not attempt to provide mechanisms for accumulating the metrics into an overall numeric evaluation. Given the different possible uses of ontologies, there is no need for such mechanisms, but rather there is a need for mechanisms capable of indicating which ontologies are more appropriate for particular situations. Also, this standard incorporates elements from the state of the art on ontology evaluation frameworks. An example of the usage of the framework is provided by evaluating two versions of the Cell Type Ontology[5]: the OBO version and a version that was re-engineered using a technique called Normalization[6].

Framework for Bio-Ontologies Quality Evaluation

In Software Engineering, software quality measures the quality of software design, and to which extent the software conforms to that design. The ISO 9126 standard for software quality evaluation provides a model based on internal, external and in-use quality metrics: functionality, reliability, portability, usability, maintainability, efficiency, effectiveness, productivity, physical security and user satisfaction. An internal metric can be used for measuring an attribute of a software product, derived from the product itself, either directly or indirectly (it is not derived from measures of the behavior of the system).Internal metrics are applicable to a non executable software product during designing and coding in early stages of the development process. An external metric can be used for measuring an attribute of a software product, derived from the

behavior of the system of which it is a part. External metrics are applicable to an executable software product during testing or operating in later stages of development and after entering to an operational process. Quality in use metrics are those applicable to the final product in real conditions.

Using such a standard as a reference for defining an ontology evaluation framework is reasonable due to the intrinsic benefits provided by the use of standards, and the context that it would provide for a systematic evaluation of ontology quality. Therefore we propose a framework for evaluating ontology quality based on such a standard. The framework comprises seven quality dimensions, and these categories have these evaluation metrics associated (Figure 1):

Structural: This category is the only one in this framework that is not specified as such in the ISO 9126, but it is important when evaluating ontologies, since it accounts for software quality factors such as consistency, formalization, redundancy or tangledness.

Functionality: How the ontology performs in its intended roles.

Reliability: Capability of an ontology to maintain its level of performance under stated conditions for a given period of time.

Usability: Readability and ease of reuse.

Efficiency: Relationship between the level of performance of the software and the amount of resources used, under stated conditions, taking into account elements such as the time response, or memory consumption. Unfortunately, the field of OE has not developed good mechanisms to evaluate efficiency appropriately.

Maintainability: The effort needed to make specified modifications, how changes affect the rest of the ontology, etc.

Quality in use: Quality in a particular context of use, provided by the users.

Next, we describe the interpretation of some of the metrics, when applied to ontologies, as follows.

Structural - Formalization: An efficient ontology has to be built on top of a semantically strict model to support reasoning. In the case of bio-ontology languages, the Web Ontology Language (OWL) has a strict semantics, the Open Biomedical Ontologies language (OBO) does not have such semantic definition, but has been defined in relation to OWL[7].

Structural - Formal relations support: Most ontologies only have formal support for taxonomy. This would indicate if any other formal theories are supporting the relations. The evaluation of this criterion for a bio-ontology depends on the number of formally supported relations included in it, for instance, through the use of the Relations Ontology (RO)[8].

Functionality - Competence adequacy- Consistent Search and Query: The formal model of the ontology allows for better querying and searching methods.

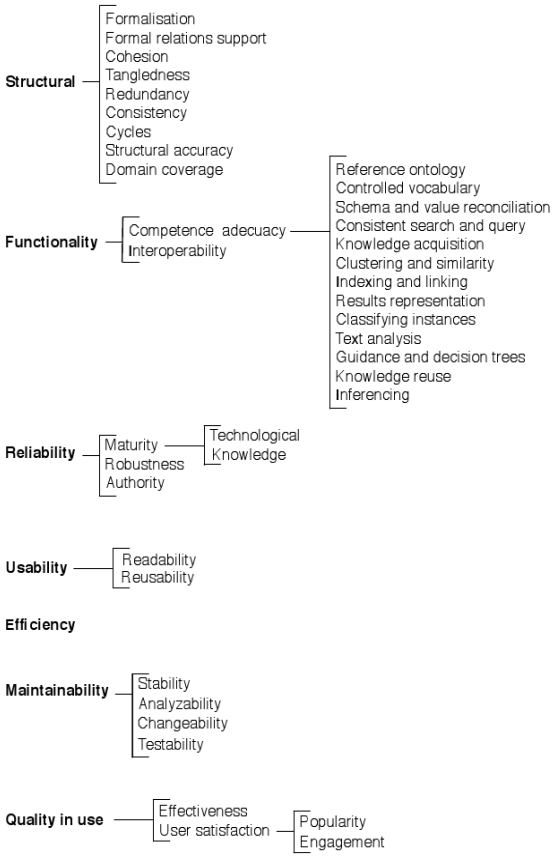

Figure 1. Evaluation framework.

Results

The Cell Type Ontology (CTO) was designed as a structured controlled vocabulary for cell types. CTO was constructed to integrate the model organism databases and other bioinformatics databases. In order to test the evaluation framework two versions of CTO were evaluated. The original version of CTO, oCTO, was the conversion of the OBO file to OWL. The normalized CTO, nCTO, was created by collaboratively dissecting the original CTO and then

recreating the structure using reasoning (http://www.gong.manchester.ac.uk/odp/html/Normalisation.html).

The evaluation of the quality of these ontologies was performed by eight MSc students of the Semantic Web course at the University of Murcia. Before doing this work, the students were trained in this course for 20 hours in the design of ontologies, they analyzed some of the most prominent ontologies (including biomedical ones) and they were also trained in the application of this evaluation framework. Then, they were given two weeks to evaluate both ontologies.

Each student had to fill in a form for each ontology, providing a quantitative evaluation for each quality metric included in the framework. The value ranged between 1(worst) and 5 (best). They were optionally allowed to provide comments on their evaluations. The usage of a quality evaluation framework does not require providing a numerical score for the evaluated items. In this case, we have averaged the results for each quality criterion for descriptive purpose, and all the quality criteria have been equally weighted. The results of this experiment are shown in Figure 2. A radar graph has been used for such purpose, since it allows an easy comparison of the quality of the two ontologies. The evaluators have given to nCTO a higher score in terms of structural, functional, usability, reliability and maintainability quality, whereas no big differences are found in terms of efficiency and quality in use.

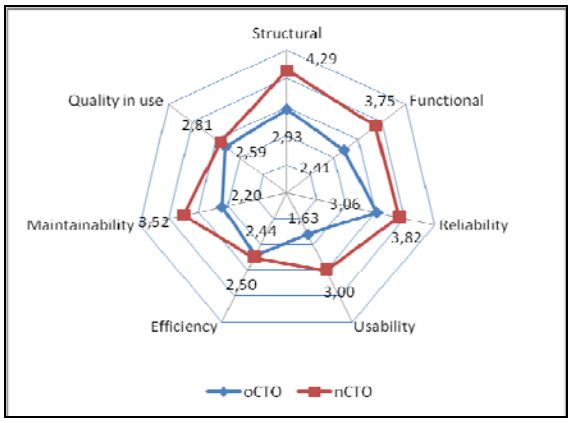

Figure 2. Results of the experiment

As has been mentioned, eight people have participated in this evaluation experiment, so the analysis of the degree of agreement between them is an interesting issue. All the evaluators gave a higher score to nCTO for the structural dimension; seven did so for functionality and usability; six did so for reliability and maintainability. It might be said that there is a consensus across these categories. In terms of efficiency, four evaluators gave a higher score to nCTO and three to oCTO. Four evaluators gave a higher value to nCTO and four to oCTO in the quality of use criterion. The evaluation of quality in use is the average of effectiveness and user satisfaction, which is split into popularity and engagement. In this sense, oCTO gets a higher score for user satisfaction, and a lower for effectiveness, due to its better structure. Hence, due to the effects of the numeric average, nCTO gets a slightly higher value for this quality dimension. So, in terms of efficiency and quality of use, there is no consensus. Both ontologies and the complete results of this experiment can be found at http://dis.um.es/~jfernand/icbo.

Discussion and Conclusions

The evaluation of ontology quality is a critical issue that remains unsolved. Different approaches accounting for different perspectives and aspects of ontology evaluation have been proposed in recent years, although none has become standard. In general, usability, reliability, and functionality criteria are identified in such approaches for evaluating quality, whereas those focused on ranking and correctness mainly consider structural properties.

In our opinion, the quality of an ontology is related to the degree of excellence. International quality organizations do not assign a numerical quality value to all kinds of processes and products, but they give them a quality stamp. This also occurs with software development processes. Such stamps certify their degree of excellence, which is checked against a series of criteria. The ISO 9126 has been criticized for being too general and abstract, and for not providing a concrete framework to be applied, obtaining a numerical evaluation as a result. The approach presented in this paper is based on the ISO 9216 and the framework includes most of the quality categories identified in the standard and incorporates the structural one to account for issues of particular importance for ontologies and it has been applied to two different ontologies, oCTO and nCTO. Both ontologies were built by applying a different methodology; oCTO was built in OBO and then transformed directly into OWL, and nCTO was built from scratch by applying the Normalization technique. This evaluation experiment has shown the usefulness of our approach, since we have obtained a vision of the quality of the ontologies, their strengths and their weaknesses, so that users have extensive information about the properties of both ontologies that can be used for making their decisions. In fact,

quality evaluation approaches do not have to make decisions for the users, but provide enough information for them to make such decisions. As mentioned, the students were trained in the evaluation framework. This training consisted on explaining the meaning of the different quality dimensions used in the framework. Examples with ontologies were provided, using good practices in ontology construction as the evaluation criteria. Obviously, the ontologies used in the training were not the ones to evaluate. Consequently, we think the scores were not biased by the training received by the students.

We were also concerned by how difficult the application of the framework could be and if this would require much technical knowledge. The students did not report problems in understanding how to apply it. This makes us think that any person with knowledge in ontology construction can do it as well without much effort. Another issue would be who should apply it and evaluate the quality of bio-ontologies[9,10], but this discussion is out of the scope of this work.It should be said that this is early work, and that some improvements are needed. This experiment is as much an evaluation of the framework as it is of the ontologies themselves. In addition, the low number of relatively inexperienced ontologists makes any profound conclusions on the nature of the two ontologies suspect. We aim to design an objective quality evaluation framework, and this has been partially achieved in this work. First, the quality dimensions and criteria are the ones defined in the ISO standard, which provides an objective definition of quality evaluation. We have added the structural dimension and defined the concrete competences of an ontology. For this, we have used standard criteria for the structural dimension, drawn from the best practices and which are generally used for evaluation purposes in literature. Concerning competences, we are using the ones considered by the community. From this perspective, the framework is objective and not biased by our interests or preferences. What is not completely objective is the measurement of the values given by the experts. We will do further research in this area to gain objectivity in this part of the process. Finally, we plan to enrich the framework including metrics related to the ontology inference power based on the theory of justification[11].

Acknowledgements

This research was funded by the Spanish Ministry of Science and Innovation through the José Castillejo Fellowship JC2008-00120, EPSRC and the University of Manchester. Work on nCTO was funded by EPSRC-funded Ontogenesis Network (EP/E021352/1).

References

1. Lopez MF, Gomez-Perez A, Sierra JP, Sierra AP. Building a chemical ontology using methontology and the ontology design environment. IEEE Intelligent Systems 1999, 14(1):37-46.
2. Aranguren ME, Antezana E, Kuiper M, Stevens R. Ontology design patterns for bio-ontologies: a case study on the cell cycle ontology. BMC bioinformatics 2008, 9(Suppl 5):S1.
3. Chua BB, Dyson LE. Applying the ISO 9126 model to the evaluation of an e-learning system. In Proceedings of the 21st ASCILITE Conference, December, 2004, Perth, Australia.
4. Al-Kilidar H, Cox K, Kitchenham B. The use and usefulness of the ISO/IEC 9126 quality standard. In Proceedings of the International Symposium on Empirical Software Engineering, November, 2005, Noosa Heads, Australia.
5. Bard J, Rhee SY, Ashburner M. An ontology for cell types. Genome Biology 2005, 6(2):R21.
6. Rector A. Modularisation of domain ontologies implemented in description logics and related formalisms including OWL. Proceedings of the 2nd International Conference on Knowledge Capture, October, 2003, Sanibel Island, USA.
7. Horrocks I. OBO flat file format syntax and semantics and mapping to OWL. Available from: http://www.cs.man.ac.uk/~horrocks/obo/.
8. Smith B, Ceusters W, Klagges B et al. Relations in biomedical ontologies. Genome Biology 2005, 6(5):R46.
9. Kalfoglou Y, Hu B. Issues with evaluating and using publicly available ontologies. Proceedings of the 4th International EON Workshop, Evaluating Ontologies for the Web, May 2006, Edinburgh, Scotland.
10. Obrst L, Hughes T, Ray S. Prospects and possibilities for ontology evaluation: the view from NCOR. Proceedings of the 4th International EON Workshop, Evaluating Ontologies for the Web, May 2006, Edinburgh, Scotland
11. Horridge M, Parsia B, Sattle U. Laconic and precise justifications in OWL. In Proceedings of the 7th International Semantic Web Conference, October, 2008, Karlsruhe, Germany.