Community-driven Ontology Evolution: Gene Ontology Case Study

Anna V. Zhdanova

ftw. Telecommunications Research Center Vienna,
Donau-City-Strasse 1/3.Stock, A-1220 Wien, Austria
zhdanova@ftw.at

Abstract. Communities on the Web capture, represent, and evolve their knowledge using ontologies, either explicitly or implicitly. The Gene Ontology project is a typical and advanced case study of a community-driven ontology creation and evolution. We use this case study to derive and illustrate factors that limit dynamic knowledge sharing in community environments. Specifically, we analyze ontology evolution implemented by the Gene Ontology community over the period of five years, as well as the used infrastructures for knowledge management. We observe limitations of communication practices within community-driven ontology construction, the lack of correlation between requested and actual changes in the ontology, and propose social and technical guidelines for making ontology-based knowledge sharing and evolution more responsive to users' needs.

Keywords: ontology evolution, collaborative semantic environments, gene ontology, community-driven ontology management

1 Introduction

Community-driven ontology construction has been applied in a number of case studies in various domains, such as in environments to support work of an organization as well as for entertainment and keeping in touch [5, 10, 16] or infrastructures for Semantic Wikis [11, 15]. The Semantic Web approach [2, 6] and development of knowledge portals [8, 13] brought numerous technologies for knowledge representation and sharing that serve as a ground for community-driven ontology evolution. Ontology evolution with social aspects [3] become generally important as there is an emerging need to understand how online communities advance their shared knowledge over time, how to measure and predict the ontology evolution rates, and how these rates co-relate with communities' successes or fruitful collaboration outcomes.

A promising area for community-driven ontology management application comprises environments supporting knowledge-intensive research communities, e.g., in eScience. Typically, life sciences can be seen as an important domain of community-driven Semantic Web application due to large amounts of domain-related information and data that needs to be exchanged between the life scientists. In particular, a charter for "Semantic Web for Health Care and Life Sciences Interest Group" (HCLSIG) has

been published by W3C[1]. Community-driven ontology construction is being addressed by the Gene Ontology (GO)[2] developers and users on a large scale. The GO Consortium [4] provides structured, controlled vocabularies and classifications that cover several domains of molecular biology and are freely available for community use in the annotation of genes, gene products, and sequences.

The GO community can be seen as far ahead of other communities in *consensus-grounded* and *collaborative* construction of ontologies [1]. Moreover, the ontology size, its high dynamics rate, years of progressive development, and large number of people involved in the project make the GO community one of the largest, representative and thus important case studies for application of community-driven semantics.

Our research objective in this paper is to observe how the collaborative community-driven ontology evolution is implemented in the real-life practical setting, and thus identify successes and further challenges for the approach. We analyze the way the GO community organizes the ontology construction process, track dynamics of GO over time periods, identify correlations between community involvement and the GO evolution, its up-to-dateness and representation. Using the GO case, we illustrate how the communities' working habits and the state of the art supporting tools can be further advanced towards making the ontology evolution driven primarily by the communities of its users and developers.

The paper is structured as follows. In Section 2, we describe the general relation between goals and usages in Web community infrastructures, and in particular the GO community: their goals and activities. Actual changes taking place in the GO over a five year period are presented in Section 3, and an analysis of these changes is provided in Section 4. In Section 5, we discuss approaches that can assist to evolve the GO. Section 6 concludes the paper.

2 Goals and activities of the community

Evidently, in many cases a community is created to reach certain goals, as it is the case for the GO community. At the same time, the reality demonstrates that once the community Web environment starts to run, it is very likely to be used to satisfy also other goals than the ones set by the community environment hosts [12]. Thus "usages" of ontologies in community environments can differ from the creation purpose of these ontologies and environments. In fact, "usages" can redefine the initial community "goals". For example, software developers may find that communities have discovered and are using added-value functionality for a certain purpose whereas the product has not initially been designed for this purpose. The developers can set catering this previously unexpected usage purposes as a goal of further development, thus demonstrating a strong connection between goals and usages.

In Figure 1, we show the main interaction levels we distinguish characterizing goals and usages in community environments, which are as follows.

[1] HCLSIG Charter: http://www.w3.org/2001/sw/hcls/charter.html
[2] The Gene Ontology: http://geneontology.org

- Considering *individual user level*: "usage", how people use the community environment,
- Considering *community level*: explicit and implicit feedback, what people say explicitly and which implicit message they express by interacting with the environment,
- Considering *community maintainers/software level*: "goals", which goals and purposes the community creators pursue when setting up community infrastructure.

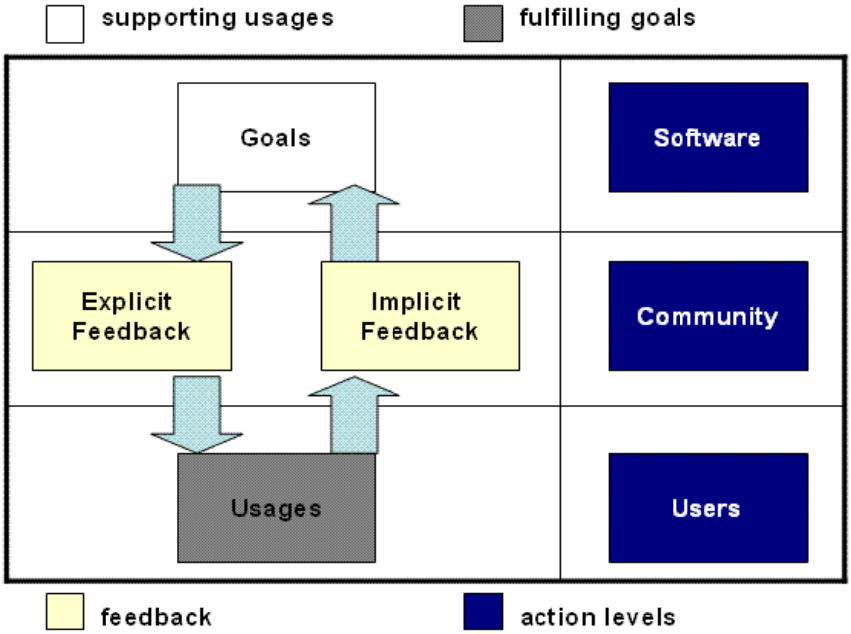

Figure 1: Goals and usages in community environments

Further, we analyze the GO community adhering to the described above principles. Specifically, the goals, feedback flows and software infrastructure of the GO community are as follows.

Main goals of the GO community are:

- collect, structure, distribute and disseminate information in the field of genomics;
- provide a set of structured vocabularies for specific biological domains that can be used to describe gene products in any organism [4].

The broader goal of Open Biomedical Ontologies (OBO) [14] is to cover the range of biology which is currently described largely in English natural language, and thus facilitate querying, analysis and "de facto" integration.

The GO community reaches its goals and performs its usages employing the following **technical infrastructure**:

- To *collect information*: mailing lists, meetings, SourceForge account;

- To *structure information*: Concurrent Versioning System (CVS), SourceForge account[3], editors such as DAG-Edit[4], formalisms such as OBO language (OBOL);
- To *distribute/disseminate information*: websites geneontology.org, sourceforge.net, CVS, converters to different ontology languages such as to OWL[5].

3 Structure and changes of GO

Structurally, GO as such consists of three sub-ontologies of the following domains:

- biological process ontology,
- cellular component ontology,
- molecular function ontology.

The molecular function defines what a gene product does at the biochemical level. The biological process normally indicates a transformation process triggered or contributed by a gene product involving multiple molecular functions. The cellular component indicates the cell structure a gene product is part of. As a whole, GO contains around 20 000 concepts.

The ontology structure is fairly simple: GO is a handcrafted ontology accepting only ”is-a” and ”part-of” relationships [14]. The hierarchical organization is represented via a directed acyclic-graph (DAG) structure similar to the representation of Web pages or hypertext systems. Members of the GO consortium group contribute to updates and revisions of the GO. The GO is maintained by editors and scientific curators who notify GO users of ontology changes via email, or at the GO site by monthly reports. When annotating the GO terms, the provided annotations should include its data provenance or source a cross database reference, a literature reference, etc.

GO is available in several different formats, such as: OBO flat file format (obo extension), GO flat file format (ontology extension), XML (RDF/XML) file format (rdf-xml extension), OWL (RDF/XML) file format (owl extension), MySQL Version.

The structure of a GO Term is as follows:

- term name (e.g., “cell”),
- a GO identifier/accession number: an arbitrary (non-semantic, meaningless) unique, zero-padded seven-digit identifier prefixed by GO (e.g., “GO:0005623”),
- optional synonyms (e.g., “synonym of apoptosis= type I programmed cell death”),
- database cross references: identifiers used to maintain cross references among databases (e.g., term “retinal isomerase activity” has the database cross reference EC:5.2.1.3 which is the accession number of this enzyme activity in the Enzyme Commission database),
- definition (e.g., ”The action characteristic of a gene product.”),

[3] The Gene Ontology Project of SourceForge: http://geneontology.sourceforge.net

[4] DAG-Edit, A controlled vocabulary editor: http://sourceforge.net/project/showfiles.php?group_id=36855

[5] Web Ontology Language: http://www.w3.org/2004/OWL

- comment (e.g., “Note that this term refers to both the old and new”).

An example of a GO term description in the OBOL 1.2 notation is as follows.

```
[Term]
id: GO:0000015
name: phosphopyruvate hydratase complex
namespace: cellular_component
def: "A multimeric enzyme complex, usually a dimer or an
octamer, that catalyzes the conversion of 2-phospho-D-
glycerate to phosphoenolpyruvate and water." [GOC:jl,
ISBN:0198506732 "Oxford Dictionary of Biochemistry and
Molecular Biology"]
subset: gosubset_prok
synonym: "enolase complex" EXACT []
is_a: GO:0043234 ! protein complex
is_a: GO:0044445 ! cytosolic part
```

All *changes in the gene ontology* are listed explicitly in monthly reports[6], apart from being steadily executed in the shared repositories runtime. The monthly reports contain a concise summary of what has happened in the GO ontologies over the past months, specifically, information about

1. new terms,
2. term name changes,
3. new definitions,
4. term merges,
5. term obsoletions,
6. significant term movements,

as well as general statistical data for the ontologies, such as total quantities of terms for every sub-ontology and the items from the SourceForge tracker that have been closed over the past month.

Addressing the issues of explicit and implicit feedback in the GO community, one has to primarily notice that *active curation* of the GO construction is one of the GO success factors [1]. GO construction is moderated by about 40 GO team members. However, involvement of a broad community of ontology users is limited to their provision of suggestions on ontology modification. Such approach to ontology construction can be seen as restrictive in the light of current consensus modeling solutions which provide community members more opportunities to be involved in ontology construction [16].

Explicit feedback (i.e., what community members request to change) is mainly performed via SourceForge. Specifically, any community member can submit a suggestion on GO modification, e.g., as a “curator request” for issues on the ontology terms. Four categories are offered to choose from when a request is submitted: “new term request”, “other term-related request”, “term obsoletion”, and “none”. Explicit feedback features from sourceforge.org have been available from February 2002 and in March 2002, the first SourceForge requests started to get resolved by the GO curators.

In Figure 2, we indicate how many “curator requests” to change the gene ontology were explicitly proposed by the community (the lower line of the graph). As for the

[6] GO community monthly reports: http://www.geneontology.org/MonthlyReports/

ontology evolution as a whole, a steady increase on the work around GO is observed, both in terms added and in relations between these terms. In fact, the number of relations between the terms grows considerably rapidly than the number of terms [9]. On the graph in Figure 2, we summarize the total number of changes in GO (the upper line of the graph) over time. The horizontal axis indicates the time period and the vertical axis indicates the total number of ontology changes.

4 Data analysis

In this section, we provide the community-driven ontology evolution challenges and limitations derived from the case study. We also discuss the publicly available case study datasets (see Section 2 and 3 for the references) and the specifics of measuring or counting the changes in an ontology developed by a community.

4.1 Challenges and limitations of community-driven ontology evolution

Analyzing the GO dynamics data, certain challenges or issues can be identified with respect to the *general community involvement* in ontology construction:

- dynamics of the ontology development only weakly correlates with the development of the actual domain (biology): in particular, at certain points in time substantially more changes are made merely because the curators are more active or because a major formal restructuring of the ontology takes place (as for example, the modification "peak" of April 2003 represented with the upper line of the graph in Figure 2);
- in some cases, actual community involvement goes beyond model organism database communities. For example, the development of the immunology component (comprising 726 new terms) of the GO did not come from any particular model organism database community, but from the immunology community [7]. However, the GO community does not have typical automatized practices for integration of the ontology construction input from adjacent communities. Such integration is performed via ad-hoc modeling and merging, and face to face meetings between the community representatives;
- certain relatively old (e.g., dated from 2002) curator requests are still marked as "open", which shows that the communication process in the community can be improved by employment of an infrastructure allowing support of alternative versions and enabling communities to agree on some parts of these ontologies.

Regarding evaluation of the *quality of community involvement* in the ontology construction and the *completeness and comprehensiveness* of the ontology, the following observations have been derived from the GO community case study:

- implicit feedback (how GO is actually used) is currently directly not considered in ontology construction;
- SourceForge requests from the community are far from directing or causing the majority of changes: as one can see from Figure 2, most of the changes done in the GO are still curator/expert-driven;

- pre-established categories of ontology changes are not equivalently important (e.g., "new terms" are introduced significantly more often than "term merges" take place). Therefore, initial (not user community-driven) categorization of the GO construction operations appears to be an ad-hoc set up. Such pre-categorization and the predefined by experts ontology cannot be comprehensive;
- the effort distribution among the development of ontology items is likely to be inadequate, as long as it is expert, and not community-biased. As the implicit feedback of the user community is not considered, more often demanded ontology items will not necessarily be the ones that are better specified and promoted. Only if new "organisms" formally join the consortium, the specification efforts are triggered. For instance, when plants joined, major ontology evolution work was needed, due to the areas in the domain totally not considered before (even though they have always been relevant): flies, mouse, yeast do not have twigs and leaves, etc.

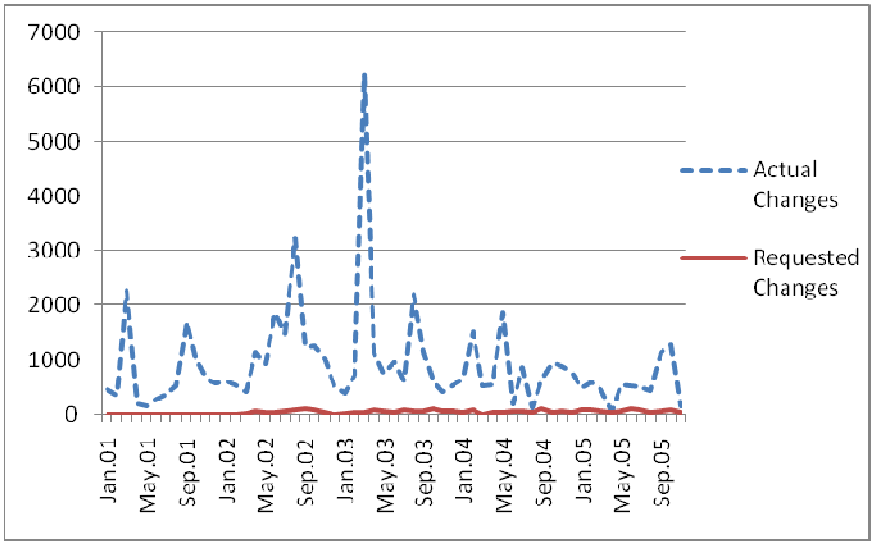

Figure 2: Total number of changes requested vs. implemented in the GO

4.2 Measuring community-driven ontology evolution

As indicated earlier, the changes occurring in the GO are classified in six categories: new terms, term name changes, new definitions, term merges, term obsoletions, significant term movements. It should be questioned whether the changes stemming from all the categories are equally important. In particular, one could notice that some changes could be considered as "bulk" modifications that result from a single request and a larger scale operation. For example, the large peak of changes occurring on March 2003 shown in Figure 2 is highly influenced by many homogeneous term name

changes in the function ontology: these term name changes constitute ca. 81% of the total changes in the whole GO for this month.

In order to diminish the importance of the "bulk" changes and generally attain a balanced representation of modification rates across different categories, we propose a normalized measure for calculating an "adjusted" weight of total changes in one category over a specific period of time. Using the adjusted weights instead of the weights calculated as the sum of the all the changes provides a semantic (versus "mechanic") view on the ontology evolution. The formula for calculation of the adjusted weight is shown in Equation 1. There, $y_{a,b}$ is a new resulting adjusted weight of the ontology changes made for month a and change category b, and $x_{a,b}$ is an absolute number of such changes (as taken as an input for Figure 2). The total number of time periods (here, months) analyzed in the use case accounts from 1 to n and is denoted as i. The total amount of ontology change categories present in the use case accounts from 1 to m and is denoted as j. The total adjusted weight of changes for a specific month can be obtained as the sum of the adjusted weights of all change categories for this month.

$$y_{a,b} = x_{a,b} \frac{1}{m} \frac{\sum_{i,j=1}^{n,m} x_{i,j}}{\sum_{i=1}^{n} x_{i,b}} \quad (1)$$

Eq. 1: Adjusted weight of ontology changes for month a and category b

The formula is designed to decrease the weights of the change categories where the large number of changes is regularly occurring on a "mechanical" or "bulk" basis and increase the weights of the changes that are performed more rarely and require more community members' inputs and attention. In particular, application of the Equation 1 for month of March 2003 and the category "term name changes", the resulting weight of the changes in this category is 4206 vs. 5052 changes in an un-adjusted weight. For the observed 5 year period, the changes in the category "term name changes" occurred more frequently than average changes in all categories by ca. 20%. The later means that in calculation of the adjusted weights, changes made in this category are counted as less significant than initial number-based weight, while changes in few other categories result in higher adjusted weights than initial weights.

The formula of Equation 1 especially well reflects the situation where the intensiveness of the community feedback needs to be correlated with the intensiveness of the actual ontology development. Different weight adjustment formulas for actual and requested changes can be constructed depending on the needs of the use cases for community driven ontology evolution. Here, as the requested changes the ontology change requests do not normally come in "bulks" but are user-generated on a one-by-one basis, the requested changes numbers are not as substantially influenced by the formula as the adjusted weights for actual changes. However, as the GO use case is not community-driven to its full potential, even after application of the formula and recalculation of the weights of the requested vs. actual changes, the discrepancy between the community-driven and the actual ontology evolution remains.

5 Towards increased community support

As described in Section 3, the explicit feedback to the ontology editing can be partially obtained via ontology development environments like SourceForge. However, current community-driven ontology evolution is hardly influenced by the communities' *implicit feedback*, and we are not aware of tools for integration of such feedback in the ontology evolution process. The contribution of implicit feedback to the ontology construction could be substantially improved by applying emerging social software practices. In particular,

- applications employing the ontology could automatically report the difficulties encountered by the GO users;
- relying on the community in adding, making obsolete or dislocating ontology terms, and not on a curator. The latter involves considerable amount of human labor, long time to perform the change, and the risk of acquiring a single point of failure in the process;
- new and existing ontology items could be automatically suggested to other parties who are potentially interested in these items. For example, as discussed in the community-driven ontology matching approach, the ontology items can be suggested for use to other people under condition that the users belong to relevant communities and social networks [17].

Here we list challenges that are needed to be overcome in the current ontology and knowledge management practices in order to attain the community-driven ontology evolution [16]. In

Table **1**, we name these challenges in its first column. The requirements on enabling technical infrastructures are listed in the second column of the table, and the requirements on communities' habits are listed in the third column of the table.

Bringing community-driven ontology evolution to the GO and similar communities is targeted at the *following audiences*:

- Developers of various community environments (to illustrate by example the influence of user and developer communities on ontology construction process, and define requirements to the infrastructures allowing benefit from its communities at the highest degree);
- Developers of tools supporting ontology evolution and versioning (to give an idea on which ontology change operations are especially useful and can be successfully captured and processed by the community);
- Computer scientists community, to spot gaps in the market with the case studies for community-driven ontology construction, such as for the GO communities.

In this paper, technical factors of community-driven ontology construction have been considered in more detail than social and organizational factors. This consideration is intentional as (i) the paper mainly addresses the readers with a technical background, (ii) collection and analysis of the social and organizational factors are very complex and go beyond ontology engineering. For instance, a specific influence of the GO community face-to-face meetings or the impacts of the GO advisory board on the ontology evolution are very difficult to capture, represent and estimate quantity-wise. Once the technical support enables to manage and measure aspects related to

social and organizational factors of the community-driven ontology evolution, the impact of such factors would need to be further analyzed.

Challenge	**Infrastructure functionality requirements**	**Community habits requirements**
Acquisition of ontologies and annotations	Enable *large scale implicit automatic production of ontologies and semantic annotations* that represent results of users' activities, such as creation and reuse of ontology items in applications, references to existing instance data, etc.	Getting used to *authorized sharing of the data arising from person's activities*; understanding of *security and privacy issues*
Notification and search of ontology items	*Retrieval of relevant ontology items* on the basis of usage histories, personal profiles and preferences, social network and community information	*Learning how to use more complex search, notification and user rules technologies* to their full potential
Ontology visualization and usability	*Visualizing ontology* taking into account importance of the item to a specific user, selection of the *usable presentation mode*	*Understanding of the subjective character of information that is seen*, i.e., why the software demonstrates the ontology and the annotations in the way it does
Versioning support	*Maintenance and usage of the versioning history* in processing of ontology items and annotations	*Getting accustomed to consider the date of the items*, e.g., when searching or designing applications that take into account versioning information
Scalable, distributed infrastructure support	*Accessing and using information from different sources; extracting knowledge from large volumes of information*	Understanding that multiple heterogeneous *information sources can vary in stability, reliability and trustworthiness*
Mobility	*Connectivity between devices* used by community members to work with knowledge; an opportunity to easily switch from one device to another	A habit to *contribute to the community from wherever possible*, in particular, situations not only limited to a typical office setting

Table 1: Habits and infrastructure requirements for community-driven ontology evolution

6 Conclusions

The paper investigates the current community-driven ontology evolution principles, taking the GO community practices as an advanced use case for ontology-based knowledge sharing and evolution. Discrepancies in correlation between the requested by the community vs. the actually performed ontology changes are identified and illustrated. The guidelines for development, measurement and use of advanced infrastructures supporting community-driven ontology evolution are suggested on the basis of the GO case study. As GO is highly community-driven, observations and conclusions drawn from its case shall serve as a flagship for numerous less developed cases of community-driven ontology evolution.

Curator-driven approach made the GO project a success of community-driven ontology development even before the later trend became common with community-driven ontology management, Semantic wikis and Web 2.0 technologies. The restrictiveness of the curator-driven approach keeps the ontology development under control to a higher degree and presumably helps to keep more facts in the ontology scientifically justified or "correct". However, together with the imperfections of the current ontology development infrastructures the "curated" approach makes the resulting ontology less up-to-date, receiving less user feedback and less complete and representative than it could be. Challenges that are yet to be addressed in the community-driven ontology evolution field are identified, and they generally shall address (i) finding a balance between the "curated" and the purely user-generated content and (ii) combining the technologies that assist to formalize and process more information about the community, including its currently implicit social and organizational factors.

References

[1] Bada, M., Stevens, R., Goble, C., Gil, Y., Ashburner, M., Blake, J., Cherry, M., Harris, M., Lewis, S., 2004. "A short study on the success of the Gene Ontology", *J. Web Sem. 1(2)*: 235-240.

[2] Berners-Lee, T., Hendler, J., Lassila, O., 2001. "The Semantic Web", *Scientific American 284(5)*, pp. 34-43.

[3] Chen, L., Haase, P., Hotho, A., Ong, E., Mauroux, P.C. (Eds.), 2007. Proceedings of an ISCW+ASWC'07 International Workshop on Emergent Semantics and Ontology Evolution (ESOE'07), November 12, 2007, Busan, Korea.

[4] The Gene Ontology Consortium, 2001. "Creating the gene ontology resource: design and implementation", *Genome Research, 11(8)*:1425–33.

[5] Corcho, O., Gomez-Perez, A., Carmen Suarez, M., 2006. "The ODESeW platform as a tool for managing EU projects: the KnowledgeWeb case study", In *Proc. of 15th International Conference on Knowledge Engineering and Knowledge Management Managing Knowledge in a World of Networks (EKAW)*, 2-6 October 2006, Podebrady, Czech Republic.

[6] Davies, J., Fensel, D., van Harmelen, F. (Eds.), 2002. Towards the Semantic Web: Ontology-Driven Knowledge Management, John Wiley & Sons.

[7] Diehl, A.D., Lee J.A., Scheuermann R.H., Blake J.A, 2007. „Ontology development for biological systems: immunology", *Bioinformatics 23(7)*:913-5.

[8] Maedche, A., Staab, S., Stojanovic, N., Studer, S., Sure, Y., 2003. "SEmantic portAL - The SEAL approach", In: Fensel, D. et al. (Eds.), *Spinning the Semantic Web*, MIT Press, Cambridge, MA, pp. 317-359.

[9] Mungall, C., 2004. Increased complexity of GO,
URL: http://www.fruitfly.org/~cjm/obol/doc/go-complexity.html.

[10] O'Murchu, I., Breslin, J.G., Decker, S., 2004. "Online Social and Business Networking Communities", In *Proc. of ECAI 2004 Workshop on Application of Semantic Web Technologies to Web Communities*.

[11] Riehle, D. (Ed.), 2005. Proceedings of the 2005 International Symposium on Wikis (WikiSym 2005), October 16-18, 2005, San Diego, California, USA.

[12] Shirky, C., 2003. "A Group is Its Own Worst Enemy: Social Structure in Social Software", *Keynote talk at the O'Reilly Emerging Technology Conference*, Santa Clara, US, April 24, 2003.

[13] Staab, S., Angele, J., Decker, S., Erdmann, M., Hotho, A., Maedche, A., Schnurr, H.-P., Studer, R., Sure, Y., 2000. "Semantic Community Web Portals", *Computer Networks 33(1-6)*, pp. 473-491.

[14] Stevens, R., Wroe, C., Lord, P., Goble, C., 2003. "Ontologies in bioinformatics", In: Staab, S., Studer, R. (Eds.), *Handbook on Ontologies in Information Systems*, pp. 635–657.

[15] Völkel, M. Krötzsch, M., Vrandecic, D., Haller, H., Studer, R., 2006. „Semantic Wikipedia", In *Proceedings of the 15th international conference on World Wide Web, WWW 2006*, Edinburgh, Scotland, May 23-26, 2006.

[16] Zhdanova, A.V., 2008. "Community-driven Ontology Construction in Social Networking Portals", *International Journal on Web Intelligence and Agent Systems, Vol. 6, No. 1*, IOS Press, to appear.

[17] Zhdanova, A.V., Shvaiko, P., 2006. "Community-Driven Ontology Matching". In *Proc. of the 3rd European Semantic Web Conference (ESWC'2006)*, 11-14 June 2006, Budva, Montenegro, Springer-Verlag, LNCS 4011, pp. 34-49.