Bridging the gap between social tagging and semantic annotation: E.D. the Entity Describer

Benjamin M Good[1§], Edward A Kawas [1], Mark D Wilkinson[1]

[1]The James Hogg iCAPTURE Centre for Cardiovascular and Pulmonary Research,
Providence Health Care/University of British Columbia,
St. Paul’s Hospital, Rm. 166, 1081 Burrard St.
Vancouver, British Columbia, V6Z 1Y6
Canada

[§]Corresponding author

Email addresses:
BMG: goodb@interchange.ubc.ca
EAK: ekawas@mrl.ubc.ca
MDW: markw@illuminae.com

Abstract

Semantic annotation enables the development of efficient computational methods for analyzing and interacting with information, thus maximizing its value. With the already substantial and constantly expanding data generation capacity of the life sciences as well as the concomitant increase in the knowledge distributed in scientific articles, new ways to produce semantic annotations of this information are crucial. While automated techniques certainly facilitate the process, manual annotation remains the gold standard in most domains. In this manuscript, we describe a prototype mass-collaborative semantic annotation system that, by distributing the annotation workload across the broad community of biomedical researchers, may help to produce the volume of meaningful annotations needed by modern biomedical science. We present E.D., the Entity Describer, a mashup of the Connotea social tagging system, an index of semantic web-accessible controlled vocabularies, and a new public RDF database for storing social semantic annotations.

Keywords

semantic web, annotation, social tagging, RDF, knowledge acquisition, mass collaboration, GreaseMonkey, mashup

Introduction: Semantic Annotation in the life sciences

Semantic annotation means the association of a data entity with an element from a classification scheme. As opposed to free-text annotation, semantic annotation is amenable to interpretation that does not require natural language understanding and hence is easily amenable to machine processing. Examples of semantic annotation include the assignment of MeSH descriptors to citations in MEDLINE and the assignment of Gene Ontology terms to gene products in UniProt [1-3].

Current practices for creating semantic annotations within the life sciences, though varying across initiatives and often augmented by automated techniques, generally follow a fairly traditional process closely resembling that long employed by practitioners in the library and information sciences [4, 5]. First, semantic structures, such as ontologies, are created by teams of life scientists working in cooperation with experts in knowledge representation (or by individuals with expertise in both areas). Next, annotation pipelines are created whereby professional annotators utilize the relevant semantic structures to describe the entities in their domain (e.g. gene products, manuscripts, micro-array experiments). Those annotations are then stored in a database that is made available to the public via websites and sometimes web services. As time goes on, the semantic structures and the annotations are updated based on feedback from the community and from the annotators themselves.

This traditional semantic annotation process yields useful results (though there are noted problems in inter-annotator agreement[6]), but it is intensive in its utilization of an inherently limited supply of professional annotators. As the technology to produce new information and the capacity to derive new knowledge from that information increases, so to must the capacity for semantic annotation. Technologies that support the semantic annotation process by partially automating it, such as genome annotation pipelines [7] and natural language indexing systems [8-10], provide important help in this regard, but manual review of automated predictions remains critical in most domains [11, 12]. Thus, there is clearly a need for an increase in the number of human annotators that parallels the increase in the amount of data. If the volume of data continues to expand at current rates, it seems that the life sciences may be heading for a rather disastrous shortage of annotators.

As the life sciences are not alone with respect to the fairly recent introduction of large volumes of relatively unstructured, but very valuable information, we may seek solutions to our problems in other domains. The World Wide Web is by far the largest and likely the least well-structured repository of information on the planet and thus provides an ideal space to observe different approaches to the annotation problem. Of the many recent developments in the evolution of the Web, one that is clearly relevant to the semantic annotation problem is the emergence of social tagging; it is from this phenomenon that the Entity Describer (E.D.) originates.

In the next section, we present social tagging as a novel and important form of Web annotation, exposing its benefits and deficiencies in comparison to other annotation processes. We then suggest a change in social tagging software that is designed to alleviate some of its weaknesses and, in the process, generate the kinds of semantic annotations required by the life sciences. The remainder of the manuscript describes a prototype implementation of the E.D. semantic tagging system, concluding with lessons learnt and a variety of suggestions for future work.

Social Tagging

Social tagging systems let their users organize personal resource collections with tags. The kinds of resources contained within them are essentially unlimited, with popular examples including Web bookmarks [13, 14], academic citations [13, 15, 16], images [17], and even personal goals [18]. These resource collections are made available to the social network of their creators and often to the general public. The tags used to organize these collections are created solely by the owner of the collection (the tagger) and can serve a variety of purposes [19]. The act of adding a resource to a social tagging collection is referred to as a 'tagging event' or simply as a 'tagging'. Tagging events are composed of a tagger, a thing tagged, a collection of applied tags, and a variety of other factors that define the context of the event (time, type of resource tagged, software used, personal purpose, etc.). Figure 1 (http://bioinfo.icapture.ubc.ca/bgood/images/EDimages/EDFigures.001.tiff) illustrates the knowledge captured in a typical tagging event in which JaneTagger tags an image retrieved from Wikipedia with the tags ‘hippocampus’, ‘image’, ‘mri’, and ‘wikipedia’.

Social tagging is intriguing for one very apparent reason, there are already an *extremely* large number of social taggers and this number is rising rapidly. In November, 2006, Del.icio.us, one of the first social tagging websites, announced that it had reached 1 million registered users [20]. On August 28, 2007, Connotea, a younger social tagging site that focuses on academics, reported 48,093 users, 470,301 taggings (bookmark posts), and 153,205 unique tags. 63% (296,785) of these taggings were identified as bibliographic citations with 39% (116,105) of these citations linked directly to PubMed identifiers (Ian Mulvany, personal communication). Figure 2 (http://bioinfo.icapture.ubc.ca/bgood/images/EDimages/EDFigures.002.tiff) charts the rapid growth of Connotea in terms of number of users, number of bookmarks and number of unique tags between May 2006 and August 2007. Though Connotea remains tiny in comparison to Del.icio.us, its impressive uptake, clear, intentional connection to the biomedical research community and powerful, open application programming interface (API) renders it an ideal candidate for exploration into the possible applications of social tagging in the life sciences.

The implications of this extensive voluntary annotation activity are numerous and have only been lightly explored. Why are so many people tagging? What useful information are they generating as a collective? How might the process be improved? Can some variant of social tagging contribute to solving the problem of the semantic annotation bottleneck in the life sciences?

Social tagging works because it provides a framework that facilitates and takes advantage of a form of *passive altruism*; whereby labor that is expended to satisfy individual needs, in this case the organization and sharing of resource collections, is passively translated into emergent, collective benefit [22]. By providing mechanisms for people to organize and share their bookmarks, images, references, etc., social tagging systems provide immediate and obvious benefits to their users. Without these individualistic incentives, it seems unlikely that these systems would achieve their current volume. These benefits do not necessarily have anything to do with the collective product of the system; however, it is that collective product that makes these systems powerful. To understand social tagging systems (and other forms of social software), it is thus necessary to analyze them at a minimum of two levels, the individual and the collective.

The main tangible, personal product of participating in social tagging is a personally organized, web-based resource collection that generates straightforward mechanisms for navigating and sharing it's contents, identifying other people with similar interests, and discovering new resources. For example, if a Connotea user tags a set of references in PubMed with tags like "heart transplantation", that user has added a new way for herself to re-find those references in the future, a simple way to share that particular set of hand-selected references with colleagues, and has identified herself to other members of the Connotea community as some one who is interested in heart transplantation. Thus, social tagging contributes to the satisfaction of personal needs in terms of both information management and social networking.

From the collective perspective, the main tangible product of a social tagging system is the database of tagging events that is generated. At a minimum, this database contains the interrelated collections of all of the tags used (sometimes called the folksonomy), all of the system users, all the resources tagged and the times at which each tagging event took place. Such databases can be mined for important information about the relationships that hold between all of these entities: what resources users found important during particular periods of time, which users are similar to each other based on shared tags and resources, which resources are similar to each other, what language the users apply to creating their tags, and, of course, which tags are used to describe each resource in the system. This last, key product of social tagging begs the question of how social tags compare to professional annotation.

Social Tagging and Professional Annotation

It is reasonable to assume that, in general, tagging actions are less consistent, less clear, less complete and thus generally less useful overall than acts of professional semantic annotation. Two likely reasons for these differences are that, in contrast to professional annotators, amateur social taggers tag for a wide variety of primarily personal purposes and come from a broad range of educational backgrounds that generally do not include formal training in classification. Furthermore, the software that they utilize is generally far less powerful for the purposes of annotation than that provided to their professional counterparts.

Some evidence in favor of these assumptions regarding the relationships between social tagging and professional annotation was provided by a short analysis of a set of 42,972 Connotea references with associated PubMed identifiers[1]. For each of these references,

[1] This analysis was performed on a dataset gathered prior to January 30, 2007

MEDLINE provided an average of 12.7 main subject descriptors while Connotea users provided only 4.8 tags. The standard deviation for the number of MeSH descriptors per reference was 5.3 while the standard deviation for the Connotea tags/reference was 7.1. Given that MEDLINE produces additional annotations in the form of, for example, modifiers and chemical indexes, it is clear that the professional annotations are far more thorough and their application more consistent than the Connotea tags. Furthermore, since MeSH descriptors are drawn from a large, detailed thesaurus that is rich in hierarchical and associative relationships, each descriptor can be used in a variety of ways that are impossible with simple tags; for example, in hierarchical browsing interfaces and in automatic query expansion.

Despite these clear weaknesses, the fact remains that social tagging systems can rapidly and inexpensively produce useful descriptions of biomedical entities. Though the annotations of the references captured in Connotea so far are semantically thin, it should not be forgotten that they were captured at the minimal cost of generating one website and one relatively straightforward database. In addition, though the overlap between Connotea and PubMed is impressive and indicative, Connotea produces annotations for many useful references that are not included in PubMed such as computer science articles that are vital for research in bioinformatics. Besides citations, it is also possible to utilize Connotea to annotate entities, such as images, genes, scientific videos etc. that are not indexed by PubMed and that may not, as yet, be annotated through any other professional curatorial process. Given its domain agnostic technology (any URI can be tagged) and its open access database (accessed via Web API), Connotea or other similar tagging services may thus be able to provide a unifying middle-layer of annotation, organically linking related resources across domains and media. However, to achieve the maximum benefit from this powerful emerging resource, it is worthwhile to attempt to address some of the deficiencies in the annotations being collected.

Linking taggers and their tags to professionally designed terminologies

It is likely not possible or even desirable to change the nature or the motivations of social taggers but it is possible to change the software that they use. When a professional annotator is presented with a new unannotated resource, they are typically provided access to a pre-defined set of terms with which to create the new semantic annotation. This terminology may sometimes lack a needed concept that they might suggest adding, but the act is generally one of selecting the appropriate concepts from an extant set rather than creating a new set of concept records each time [5]. In social tagging applications, the task of terminology creation is typically left to the user of the system. At most, these applications provide access to previously used tags via a "type-ahead" window that is often limited to display only those tags already authored by the user. If MEDLINE or Gene Ontology indexers had to use such an interface, the quality of their semantic annotations would likely decrease as well.

This comparative poverty in the software provided for creating social annotations was one of the main motivations for creating E.D. E.D. thus provides an enhancement to the Connotea tagging interface that seeks to improve the utility of social tagging events by providing taggers with direct access to pre-defined, Semantic Web-accessible terminologies. This enhancement has important consequences at the moment of tagging, by providing lists of unambiguous professionally created terms to select as tags, and in the data captured from tagging events, by allowing users to intentionally form links between tags and terms from established terminologies. E.D. thus enables a new form of *social semantic annotation* in which annotations can be created by anyone, as in social tagging, but are constructed using terms from existing semantic structures, as in professional annotation.

In addition, since the terminologies that E.D. utilizes are drawn only from those resources that are accessible via the Web using widely accepted standards for knowledge representation [23, 24], each semantic tagging act can be captured and shared such that any program designed to take advantage of these standards, such as the Tabulator [25], can utilize the data automatically. If successful, E.D. will result in better personal and aggregate resource collections that, in the long term, may be used to provide powerful enhancements to the work of professional life science annotators.

Growing the E.D. knowledge base

E.D. consists of two GreaseMonkey user-scripts [26], a Jena RDF database [27], and a set of Java servlets that provide the scripts access to that database [28]. The first user-script augments the "Add to Connotea" web page that is used to post taggings to the Connotea database by providing the user access to terms from controlled vocabularies and by posting the results of their semantic taggings to the E.D. database. The second user-script utilizes this database to enhance the Connotea library pages for E.D. users by adding the ability to filter tags by source vocabulary and by implementing simple query expansions over the semantic tags. The semantic tagging database captures information about the terminologies used by E.D. users, the user's personal semantic tagging palette configuration, and about the tagging events.

Social taggers must meet the following requirements before E.D. can be used to generate semantic tags: 1) install Firefox, 2) install GreaseMonkey, 3) sign up for a free Connotea user account, 4) install the “Add to Connotea” bookmarklet, 5) install the E.D. user-scripts. With these requirements in place, a tagger who browses to or highlights a URI resource in the FireFox browser and clicks on the “Add to Connotea” bookmark will be greeted with the E.D.-enhanced Connotea web page.

Figure 3 (http://bioinfo.icapture.ubc.ca/bgood/images/EDimages/EDFigures.003.tiff) displays the unmodified tagging interface provided by Connotea. Figure 4 (http://bioinfo.icapture.ubc.ca/bgood/images/EDimages/EDFigures.004.tiff) shows the first modifications to the tagging interface as they appear when it is first visited. By selecting the new “Add Vocab” link, taggers are presented with a list of terminologies to choose from as displayed in Figure 5 (http://bioinfo.icapture.ubc.ca/bgood/images/EDimages/EDFigures.005.tiff). Clicking on the “Add Controlled Vocabulary” saves the selected vocabulary in a personal list and brings the user back to the “Add to Connotea” page which now contains the newly added vocabulary in the users tagging palette. Figure 6 (http://bioinfo.icapture.ubc.ca/bgood/images/EDimages/EDFigures.006.tiff) displays a tagging palette containing six, color-coded, controlled vocabularies. In Figure 6 (http://bioinfo.icapture.ubc.ca/bgood/images/EDimages/EDFigures.006.tiff), the MeSH, Brenda (br), GO biological process (bp), and Mythical Creatures (demo) terminologies are active and the user has typed 'hip' into the tagging form. The tags in the tag suggestions box thus contain the tags 'hip_br', 'hippocampus_development_bp', 'hippocampus_br', 'hippocampus_mesh', 'hippocampus_myth', and 'hippocastanacea_mesh'. The definition shown resulted from mousing-over the hippocampus_mesh tag and corresponds to a definition gleaned automatically from the MeSH terminology.

After an E.D. user has used the interface presented in figure 6 (http://bioinfo.icapture.ubc.ca/bgood/images/EDimages/EDFigures.006.tiff) to enter all of the semantic tags and/or regular tags that they desire, clicking the "Add to Library" button sends all of the tags for that bookmark to Connotea as usual; however, they are also sent to the E.D. semantic tagging database. The vital additional knowledge captured by E.D., as a supplement to Connotea, is the connection between the newly created tags and the URIs for the controlled terms selected by the tagger. Figure 7 (http://bioinfo.icapture.ubc.ca/bgood/images/EDimages/EDFigures.007.tiff) illustrates how the social semantic tags provided by E.D. relate to normal tags and to terms from controlled vocabularies. To be clear, when an E.D. user uses a term from a controlled vocabulary as a tag, the term is represented both in Connotea and in E.D. as an independent tag that is separate, but related to the source term. This is done to ensure that no important knowledge is lost relating the user, the tag, and the tagging event by capturing this information in records associated with the new tag.

Knowledge base structure

E.D. is designed to be as lightweight as possible. It does not re-represent terms from the extant terminologies that it takes advantage of. Thus, it must directly utilize the representational structures present in the input vocabularies. For the current prototype, the primary source of OWL/RDF controlled vocabularies is the collection available from [29]. As such, the focus was placed on utilizing the semantic structure of those terminologies. At present, these resources are organized as follows:

1. Each concept from a chosen terminology is represented as an OWL class with a unique, semantically opaque identifier, a human readable label, and often a human readable description.
2. The terminologies contain a mixture of modeling approaches including, for example, the gene ontology with its is-a and part-of relations and MeSH with its broader-than / narrower-than organization.
3. In the current[2] OWL instantiations of these terminologies, the narrower-than relations and the is-a relations are both represented with the RDF-S:subClassOf property.

[2] These ontologies are primarily experimental, automatically generated, and subject to retirement when the NCBO [30] National Center Biomedical Ontology. Available at http://bioontology.org/ subsumes the functionality of this collection.

As a result of the structure of these primary inputs, the current version of ED utilizes only the relations: RDF-S:Label (to get the string for the tag), RDF-S:Comment (to display a comments), oboInOwl:Definition (to display definitions), and RDF-S:subClassOf (to represent all hierarchical relations).

E.D. must bind the tags it generates to the terms from the controlled vocabularies, in this case the represented as RDF-S:Classes, if it is to take advantage of the structure of these terminologies. This is accomplished through the use of the RDF-S:isDefinedBy predicate (see Figure 7 http://bioinfo.icapture.ubc.ca/bgood/images/EDimages/EDFigures.007.tiff). The isDefinedBy predicate is an annotation property (a sub-property of rdfs:seeAlso); thus it can be used to relate any two RDF resources, including classes, without any positive or negative consequences when the knowledge base is submitted to a reasoner. In contrast, if a non-annotation property and was used to relate a URI resource and an RDF-S:Class, it would render the knowledge base OWL-Full[3] and would thus decrease the availability and predictability of suitable reasoners such as Pellet (though at present no reasoning is done on the ED database itself and only RDF-S reasoning is implemented for the imported ontologies) [31].

An immediate and natural extension to E.D. will be support for terminologies represented using the SKOS ontology because it most correctly models the thesaurus-based structures utilized by critical terminology resources such as MeSH [32].

[3] OWL is divided into three main branches, increasing in expressivity from OWL-Lite to OWL-DL, to OWL-Full. Provably efficient reasoning algorithms have thus far only been implemented up to OWL-DL.

Using the E.D. knowledge base within Connotea

The E.D. knowledge base is used by the second GreaseMonkey user-script to provide an enhanced view over the Connotea reference collections owned by E.D. users. After installation, the tags on a user library page may be filtered based on their source vocabulary and the terms may be expanded up their hierarchical relations for browsing and down these relations for query. Figures 8 (http://bioinfo.icapture.ubc.ca/bgood/images/EDimages/EDFigures.008.tiff) and 9 (http://bioinfo.icapture.ubc.ca/bgood/images/EDimages/EDFigures.009.tiff) illustrate an example where an E.D. user has tagged a reference with the term 'hippocampus' from the MeSH vocabulary. Figure 8 (http://bioinfo.icapture.ubc.ca/bgood/images/EDimages/EDFigures.008.tiff) illustrates how that user's library page would look when the tags were filtered to only include those originating from MeSH and with expansion turned off. Figure 9 (http://bioinfo.icapture.ubc.ca/bgood/images/EDimages/EDFigures.009.tiff) shows what the page looks like with expansion turned on and the letters 'bra' typed in the 'find tag' query box'. Upwards expansion facilitates browsing through tag collections by producing additional terms like 'brain' and 'limbic system' even when they are never used directly as tags. Clicking on these broader level tags then executes a query for narrower terms with the effect that clicking on 'brain' will return references tagged with both 'brain' and narrower terms such as 'hippocampus'.

Discussion

E.D. successfully provides one of the first means of directly connecting the Semantic Web with the Social Web in a biomedical context. By creating a mechanism for social taggers to intentionally form connections between their tags and concepts from controlled, structured terminologies, E.D. makes it possible for the broad community of biomedical researchers to provide semantic annotation for any URI and have that annotation stored in a public, queriable RDF database. If the widespread and growing use of social tagging systems is any indication, the E.D. knowledge base could rapidly become a major repository of social semantic annotation for life science resources.

In the unlikely event that the product of this social semantic annotation process were as thorough, consistent and clear as the products of professional annotation, this knowledge base alone would go far towards solving the problem of the semantic annotation bottleneck. However, because of the fundamental qualitative differences between social tagging and professional indexing [33], E.D. and its descendents will likely always produce diverse kinds and qualities of annotations, that, when compared directly to professional annotation, will appear error prone, idiosyncratic, and inconsistent. Successful application of the products of E.D. will therefore depend on application and extension of emerging techniques that both utilize and compensate for the large volume of heterogeneous-quality data gathered by social tagging applications [34-36]. Given expected improvements and appropriate adaptations of these techniques, social semantic tagging initiatives such as E.D. may fulfill a key role in future approaches to semantic annotation in the life sciences, filling the gap between professional curation and automatic indexing.

Future directions

The E.D. project and the social tagging phenomenon in general are still producing substantially more questions than answers. Here we list just a few of the possible immediate modifications to the E.D. annotation system. Aside from exploring these potential improvements, future work will investigate the data gathered by E.D. for trends that may be utilized to both characterize the nature of the interaction between social taggers and pre-defined terminologies and to identify solid approaches to automatically inferring levels of trust for social annotations.

1. A key to the future development of E.D. is the incorporation of large, established terminology resources such as WordNet, the National Cancer Institute Thesaurus, the Foundational Model of Anatomy and the Unified Medical Language System [37-40]. In this first prototype, the focus was to utilize terminologies accessible via currently operational Semantic Web technology. This meant that the terminologies had to be expressed in OWL/RDF and that they had to be reasonably small. Optimizations to our code as well as to the general purpose Semantic Web stack will help expand E.D.'s capabilities in regard to supporting very large terminologies. In addition, it is possible and desirable, for E.D. to directly access web services provided directly by the terminology developers rather than bringing them into its local context and serving them itself as it does now.

2. One of the major problems faced by E.D. is the unfamiliarity of users with the terminologies. Better support for end-user interaction is already on the agenda of leading ontology development groups [41] and E.D. will clearly benefit from progress in this direction.
3. Though a type-ahead interface is clearly useful when users have authored the terms themselves and thus have some idea what is available, it is unclear if it is the best possible way to engage them with terms from vocabularies that they may be unfamiliar with. It would be valuable to execute user studies to identify the benefits and weaknesses of the type-ahead in this context and to suggest other modes of interaction; for example, it may be beneficial to display aspects of the term hierarchies via a tree-like interface or frequency of use via a tag-cloud style interface.
4. Another approach that may benefit the user-interface experience is to provide suggestions for terms either already used as indexes for the resource (e.g. MeSH for biomedical documents) or, when the resource contains textual content, through the application of text mining techniques, such as the MetaMap program from the UMLS [42].
5. Like all public social tagging applications that we are aware of, E.D. only supports the creation of annotations along the rather ambiguous "hasTag" predicate. Resource X hasTag tag Y. In the future, we would like to investigate the possibility of letting users add another level of clarity to their personal web annotations through the use of typed predicates. For example, [Resource X has part Z], [X has author Q], [X develops from W], [X is the same as I], etc.

6. We would like to extend E.D. so that it could easily be applied in other social tagging applications such as CiteULike, Del.icio.us, and Bibsonomy. The E.D. database is agnostic to the tagging service utilized, thus only the tagging interface itself would have to be modified. We are also interested in the possibility of creating a standalone service that would offer the potential for a richer, more stable user-interface by removing the constraints imposed by the current user-script approach. The database underlying such a distributed set of services might form the foundation for a unified, large-scale semantic annotation repository of widespread utility.

Conclusions

The social semantic annotation methodology embodied by E.D. should be applicable to any domain that has produced controlled terminologies and published them using OWL/RDF. Given the recognition of the importance of semantic annotation and the resultant accumulation of many structured terminology resources as well as the large and enthusiastic user community, the biomedical sciences provide an ideal application environment for E.D.; however, the core methodology may also be used in other domains that face similar challenges. While the E.D. methodology can clearly accumulate social semantic annotations, how those annotations might best be applied remains an open question.

Availability

The E.D. user-scripts and instructions for their installation and use are freely available from http://www.connotea.org/wiki/EntityDescriber.

Acknowledgements

MDW and EAK are funded through an award from Genome Prairie and Genome Alberta, in part through Genome Canada, a not-for-profit corporation leading Canada's national strategy on genomics. MDW is also supported by an award to the iCAPTURE Centre from the Michael Smith Foundation for Health Research. BMG is supported by an award to the Better Biomarkers in Transplantation project from Genome British Columbia, in part through Genome Canada. Core funding for the laboratory and related research activities has been granted from the Canadian Institute for Health Research, the Natural Sciences and Engineering Research Council of Canada, and the Heart and Stroke Foundation for BC and Yukon.

References

[1] The Basics of Medical Subject Headings (MeSH). Available at http://www.nlm.nih.gov/bsd/disted/mesh/index.html

[2] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics. 2000;25: 25-9.

[3] Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS. UniProt: the Universal Protein knowledgebase. Nucleic Acids Res 2004;32: D115 - 119.

[4] Bachrach C, Charen T. Selection of MEDLINE contents, the development of its thesaurus, and the indexing process. Medical Informatics 1978;3: 237-254.

[5] Use of MeSH in Indexing. Available at http://www.nlm.nih.gov/mesh/intro_indexing2007.html

[6] Camon E, Barrell D, Dimmer E, Lee V, Magrane M, Maslen J, Binns D, Apweiler R. An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. BMC Bioinformatics 2005;6: S17.

[7] Hubbard T, Aken B, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer S, Fitzgerald S, Fernandez-Banet J, Graf S, Haider S, Hammond M, Herrero J, Holland R, Howe K, Johnson N, Kahari A, Keefe D,

Kokocinski F, Kulesha E, Lawson D, Longden I, Melsopp C, Megy K, Meidl P, Ouverdin B, Parker A, Prlic A, Rice S, Rios D, Schuster M, Sealy I, Severin J, Slater G, Smedley D, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wood M, Cox T, Curwen V, Durbin R, Fernandez-Suarez X, Flicek P, Kasprzyk A, Proctor G, Searle S, Smith J, Ureta-Vidal A, Birney E. Ensembl 2007. Nucleic acids research 2007;35: 610-617.

[8] Aronson A, Bodenreider O, Chang H, Humphrey S, Mork J, Nelson S, Rindflesch T, Wilbur W. The NLM Indexing Initiative. Proceedings / AMIA ... Annual Symposium. AMIA Symposium 2000: 17-21.

[9] Ruch P. Automatic assignment of biomedical categories: toward a generic approach. Bioinformatics 2006;22: 658-658.

[10] Kim W, Aronson A, Wilbur W. Automatic MeSH term assignment and quality assessment. Proceedings / AMIA ... Annual Symposium. AMIA Symposium 2001: 319-323.

[11] Gattiker A, Michoud K, Rivoire C, Auchincloss A, Coudert E, Lima T, Kersey P, Pagni M, Sigrist C, Lachaize C, Veuthey A, Gasteiger E, Bairoch A. Automated annotation of microbial proteomes in SWISS-PROT. Computational biology and chemistry 2003;27: 49-58.

[12] Kasukawa T, Furuno M, Nikaido I, Bono H, Hume D, Bult C, Hill D, Baldarelli R, Gough J, Kanapin A, Matsuda H, Schriml L, Hayashizaki Y, Okazaki Y, Quackenbush J. Development and Evaluation of an Automated Annotation Pipeline and cDNA Annotation System. Genome Research 2003;13: 1542-1551.

[13] Connotea. Available at http://www.connotea.org/

[14] Del.icio.us. Available at http://del.icio.us/

[15] CiteULike. Available at http://www.citeulike.org/

[16] BibSonomy. Available at http://www.bibsonomy.org/

[17] Flickr. Available at http://www.flickr.com/

[18] 43Things. Available at http://www.43things.com/

[19] Golder SA, Huberman BA. Usage patterns of collaborative tagging systems. Journal of Information Science 2006;32: 198-208.

[20] del.icio.us blog entry. Available at http://blog.del.icio.us/blog/2006/09/million.html

[21] Mulvany I. In: Personal Communication; 2007.

[22] Bokardo: The Del.icio.us Lesson. Available at http://bokardo.com/archives/the-delicious-lesson/

[23] RDF Primer. Available at http://www.w3.org/TR/rdf-primer/

[24] OWL Web Ontology Language Overview. Available at http://www.w3.org/TR/owl-features/

[25] Tim Berners-Lee YC, Lydia Chilton, Dan Connolly, Ruth Dhanaraj, James Hollenbach, Adam Lerer, Danvid Sheets. Tabulator: Exploring and Analyzing linked data on the Semantic Web. In: The 3rd International Semantic Web User Interaction Workshop. Athens, Georgia; 2006.

[26] Good B, Kawas E, Kuo B, Wilkinson M. iHOPerator: User-scripting a personalized bioinformatics Web, starting with the iHOP website. BMC Bioinformatics 2006;7: 534.

[27] Jena Semantic Web Framework - Documentation Overview. Available at http://jena.sourceforge.net/documentation.html

[28] Joseki. Available at http://www.joseki.org/

[29] OBO export page. Available at http://www.fruitfly.org/~cjm/obo-download/index.html

[30] National Center Biomedical Ontology. Available at http://bioontology.org/

[31] Pellet: An Open Source OWL-DL Reasoner in Java. Available at http://pellet.owldl.com/

[32] SKOS Core Guide. Available at http://www.w3.org/TR/swbp-skos-core-guide/

[33] Tennis JT. Social Tagging and the Next Steps for Indexing. In: Jonathan Furner JTT, editor. 17th ASIS&T SIG/CR Classification Research Workshop. Austin, Texas; 2006.

[34] Wu X, Zhang L, Wu Y. Exploring social annotations for the semantic web. In: World Wide Web Conference; 2006. p. 417-426.

[35] Millen DR, Feinberg J. Using social tagging to improve social navigation. Workshop on the Social Navigation and Community based Adaptation Technologies 2006.

[36] Mika P. Ontologies Are Us: A Unified Model of Social Networks and Semantics. In: International Semantic Web Conference; 2005.

[37] WordNet. Available at http://wordnet.princeton.edu/

[38] Rosse C, Mejino JL. A reference ontology for bioinformatics: the foundational model of anatomy. The Journal of Biomedical Informatics 2003;36: 478-500.

[39] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucl. Acids Res. 2004;32: D267-270.

[40] Shah N, Rubin D, Espinosa I, Montgomery K, Musen M. Annotation and query of tissue microarray data using the NCI Thesaurus. BMC Bioinformatics 2007;8: 296-296.

[41] Gibson A, Wolstencroft K, Stevens R. Promotion of Ontological Comprehension: Exposing Terms and Metadata with Web2.0. In: World Wide Web Conference; 2007.

[42] Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: American Medical Informatics Association; 2001. p. 17-21.