Journal of Biomedical Informatics 46 (2013) 1–4

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Commentary

State of the art and open challenges in community-driven knowledge curation

Tudor Groza [a,*], Tania Tudorache [b], Michel Dumontier [c]

[a] School of ITEE, The University of Queensland, Australia
[b] Stanford Center for Biomedical Informatics Research, Stanford University, USA
[c] Department of Biology, Institute of Biochemistry and School of Computer Science, Carleton University, Canada

1. Introduction

In the past decade, ontologies have become central to the construction of intelligent decision-support systems, simulation systems, information-retrieval systems, and natural-language systems [1,2]. With the adoption of ontologies, especially by the broad biomedical community, the further development of ontologies and knowledge bases has evolved into a community-driven process [3]. This resulted in an increased number of knowledge bases published openly on the Web. Ontologies and knowledge bases are now authored and curated by more domain and knowledge experts than ever before. To ensure a high quality of the community-generated content, a well-defined curation process has to become a prominent and integral part of the life cycle of biomedical knowledge artefacts.

Several large biomedical projects are trying to apply the "wisdom of the crowds" model for building and curating their knowledge content. This model is already familiar to most experts (e.g., Web 2.0) and has already proved successful in large community projects in other domains (e.g., the Open Source movement). The emergence of different types of collaborative environments, such as Wikis, content management systems, and collaborative ontology editors, enables novel ways of curating knowledge, hence transforming the workflow from being curator-centred to being community-driven. Such systems provide the means for communities of experts in different fields to create, share and reuse knowledge collaboratively. Their goal is to foster long term expansion and maximisation of knowledge curation, extraction and reasoning, by creating live knowledge bases within their specific domains.

Based on the immediate goal, there are, in principle, two major types of systems: (i) knowledge curation platforms that focus on externalising knowledge for human consumption and where ontologies are either the foundational support or a byproduct of the curation process; and (ii) ontology curation environments focused explicitly on the collaborative creation and refinement of ontologies, and thus on crystallising knowledge for machine consumption. Independently of the type of platform, there are a series of critical aspects that prevail and that are, usually, the source of the most complex challenges. These aspects include: knowledge engineering – i.e., designing a (logical) scaffolding to be used by the platform for population and later inference and/or discovery; knowledge capturing – i.e., integrating the acquisition process into an organic workflow that enables experts to focus on the actual knowledge externalisation rather than on the elements of the process; and knowledge evolution – perhaps the most critical aspect, as it requires the decomposition of knowledge into atomic units and the management of these units from different, often interwoven, perspectives, such as concurrent and transactional acquisition or logical consistency.

In the following sections, we discuss a series of existing approaches in the two major categories we have identified above, knowledge curation and ontology curation; we show how the two accompanying papers in this special journal section contribute to the advancement of the state of the art; and finally we discuss the open issues and challenges in the community-driven curation of biomedical knowledge.

1.1. Knowledge curation platforms

The goal of typical knowledge curation platform is to enable researchers and experts in a particular field to define, detail and explore the knowledge within that field via a quality-driven collaborative curation process. In most cases, this process is similar to academic publishing by featuring peer reviews under the supervision of an editorial board. From a technical perspective, there are two types of foundations used to develop and deploy such platforms: (i) collaborative knowledge repositories; and (ii) Wikis, or Wiki-like systems.

The first category features a wealth of prominent projects in the biomedical domain, ranging from a focus on basic molecular aspects to complex interactions, reactions or relations. Basic molecular aspects include, protein and gene curation – UniProt (http://www.uniprot.org/) [4] and the Gene Ontology project (http://www.geneontology.org/) [5], gene mutations and DNA variation curation – LOVD (http://www.lovd.nl/) [6], or chemical structure curation – ChemSpider (http://www.chemspider.com/). More complex interactions, reactions or relations include pathway annotation and curation – Reactome (http://www.reactome.org/) [7] and KEGG (http://www.genome.jp/kegg/) [8].

* Corresponding author.
E-mail addresses: tudor.groza@uq.edu.au (T. Groza), tudorache@stanford.edu (T. Tudorache), Michel_Dumontier@carleton.ca (M. Dumontier).

1532-0464/$ - see front matter
http://dx.doi.org/10.1016/j.jbi.2012.11.007

In the second category, Wikis or Wiki-like systems, some of the most representative exponents are the Online Mendelian Inheritance in Man (OMIM – http://omim.org/) knowledge base [9] – focused on mendelian disorders, human genes and genetic phenotypes, GeneReviews [10] – focused on genetic disorders, and the GeneWiki suite of projects (http://en.wikipedia.org/wiki/Gene_Wiki) – focused on genes, while ArrayWiki (http://arraywiki.bme.gatech.edu/) focuses on the community-driven collection of metadata for micro-array meta-experiments from multiple primary sources [11].

The Open Source Drug Discovery (OSDD – http://www.osdd.net/) community took a different approach to support the discovery of novel therapies for neglected tropical diseases being inspired by the open source movement in Information Technology [12]. Crowdsourcing and social networking are engaging interdisciplinary scientists to collaborate on the SysBorg collaborative platform (http://sysborg2.osdd.net). SysBorg uses the Liferay infrastructure (http://www.liferay.com/) as a backbone and Semantic Web technologies to represent and share knowledge. The platform provides rich collaboration features such as: social networkings, forums, blogs, collaborative spaces, open project spaces, open laboratory notebooks, and workflow support. One very successful initiative of OSDD is the Interactome/Pathway (IPW) Annotation project [13] that has synthesised crowdsourcing and social networking methods to generate the first and largest manually curated interactome of Mycobacterium tuberculosis.

While the knowledge created and presented via these platforms is extremely valuable, its externalisation in a way that enables automatic processing or reasoning is still a work in progress. Knowledge captured in some of the repositories listed above (e.g., UniProt or ChemSpider) is now available as Linked Open Data [14], or as ontologies published using different logical formalisms (e.g., the Gene Ontology). However, these, in addition to OMIM which is available in a machine-understandable format in the context of the Human Phenotype Ontology (HPO) project [15], represent pioneering examples and, as such, knowledge engineering and publishing persists as an open research challenge. This challenge is continuously addressed, in particular, by the W3C Health Care and Life Sciences Interest Group (W3C HCLS – http://www.w3.org/blog/hcls/) and by the Open Biological and Biomedical Ontologies Foundry (OBO Foundry – http://www.obofoundry.org/) [16].

On the other hand, the literature also contains a few examples of platforms that have been developed with the dual goal of both crystallising the knowledge within a domain, as well as externalising it for automated processing. For example, the AlzSWAN platform [17] aims to model and capture the argumentative discourse (i.e., hypotheses, claims and arguments) associated with domain-specific knowledge about Alzheimers Disease. The curation process results in a series of annotations shared and discussed by experts, but also implicitly available as instances of the SWAN ontology [18]. These instances can then be augmented with domain specific concepts defined by different other ontologies, hence enabling the automatic creation of argumentative discourse networks spanning multiple publications or information sources.

Another example of such a platform is SKELETOME [19], which follows the same aims as OMIM but concentrates only on bone dysplasias. A skeletal dysplasia ontology [20] has, firstly, been developed and then used to generate the collaborative knowledge curation platform – as opposed to the other examples discussed above that usually start from an empty foundation. Subsequently, the platform uses a controlled (editorial) workflow to ensure a proper alignment between the knowledge curated by the experts and the underlying ontology, thus providing an iterative knowledge evolution cycle.

1.2. Ontology curation systems

Unlike the knowledge curation platforms, collaborative ontology curation systems focus on providing an environment in which experts can externalise and formalise the knowledge captured within a domain. The key aspect in this context is the formalisation process, as the result will always be an ontology (or a refined increment of an existing ontology). This formalisation process introduces, subject to underlying goals, several challenges that are usually absent in knowledge curation platforms, two of the most common being completeness and consistency checking (i.e., ensuring structural and logical correctness).

Although the ontology curation process is intrinsically collaborative (ontology *is a* **shared** conceptualisation of a domain [21]), tool support with real collaboration features has only recently been developed. Most of the existing work has been done by extending typical Wiki platforms (usually MediaWiki – http://www.mediawiki.org/) with semantic capabilities, hence creating Semantic Wikis. Representative examples in this category are BOWiki (http://bowiki.net/) [22], IkeWiki [23] or MoKi (https://moki.fbk.eu) [24]. BoWiki is an ontology-based data annotation and integration Wiki for the biological domain. It has been designed to annotate biological data in conjunction with a core bio-ontology (e.g., GFO-Bio [25] or BioTop [26]) and to enable biologists to create precise relations among biological entities. Consequently, the focus in on populating specific models with ontological instances and linking these instances via meaningful relations. BOWiki extends the features provided by MediaWiki with a set of semantic elements. These features include importing and re-using multiple bio-ontologies, semantic search and retrieval and consistency checking via an ontology reasoner. The other systems, i.e., IkeWiki and MoKi, provide similar functionalities, however their goal is more generic, i.e., to enable a complete knowledge engineering workflow. Consequently, the ontology building support is richer and comprises additional elements, such as defining and altering classes and their relationships. Two unique features of MoKi are its support for the collaborative and graphical editing of formal process models that has been applied in the collaborative encoding of cancer treatment protocols [27], and its mixing of unstructured information with a semi-formal and formal representation of the same content, which enables the participation of users with different expertise on the same platform, as well as the verification of the formal representation against the unstructured content.

Semantic Wikis deliver successfully the mechanisms required by collaborative ontology development. However, they are usually accompanied by a series of limitations, which range from the lack of certain functionalities (e.g., editing of advanced relations between concepts) to a fairly high entry barrier (i.e., users require Semantic Web or ontology engineering background to be able to take full advantage of the platforms capabilities). In order to deal with these issues, WebProtege (http://protegewiki.stanford.edu/wiki/WebProtege) [28] provides an organic environment that integrates content creation and online collaboration within the ontology development process itself. While relying on the full support offered by the original Protege infrastructure (http://protege.stanford.edu/) [29], WebProtege lowers the entry barrier to ontology development by using a highly customisable and pluggable user interface that can be adapted to any level of user expertise. Its extensive support for collaboration includes a series of critical features, such as threaded discussions integrated in the ontology editing environment, change management and evolution support. This comprehensive and versatile infrastructure has enabled its deployment in production settings for several large real-world projects, including the curation of the 11th revision of the International Classification of Diseases (ICD-11) led by the World Health Organisation (WHO) [30].

In parallel to the efforts discussed so far, there are other platforms that focus on a series of specific curation challenges that span the boundaries among multiple related ontologies. An example of such platform is BioPortal (http://bioportal.bioontology.org/) [31]. BioPortal is an open repository of over 300 biomedical ontologies and terminologies that offers its content not only through a Web interface, but also programmatically, through RESTFul Web services (http://www.bioontology.org/wiki/index.php/NCBO_REST_services). BioPortal acts not only as a comprehensive library of biomedical ontologies, but also as a curation platform. Users are able to add notes and discussions to classes in the ontology, to propose structured changes, and to review ontologies in the context of their own project. BioPortal also stores multiple versions of an ontology providing support for the evolution of ontologies. In addition to acting as an ontology repository, BioPortal allows users in the community to create mappings between concepts defined in different ontologies and to attach notes to them. Given the continuously increasing number of ontologies, this aspect should be an indispensable element in the knowledge engineering process, thereby ensuring a harmonious integration of the resulting ontology into the global knowledge ecosystem.

2. Special section articles

The two articles presented in this special journal section touch on two important challenges of collaborative knowledge curation: one looks at the foundational issue of ontology diffs in the context of the general evolution process, while the other focuses on social-collaboration aspects emerging from community-driven ontology building.

Malone and Stevens [32] measure the level of activity in ontology building projects, with the goal of both providing an analytical view over the collaboration process and making predictions on their future directions. They analyse several factors, ranging from basic change operations (add/delete/update) or frequency of releases to person-centric participation metrics, to lay the foundation of an overarching activity metric that could be adopted as an assessment criterion for community ontology development. From a quantitative perspective, the results of their study measure the usefulness of these factors, while from a qualitative perspective, they show a constant trend in collaborative development of ontologies across multiple communities.

Hartung et al. [33], on the other hand, concentrate on one of the foundational operations that govern ontology evolution – i.e., ontology diffs. They propose a novel approach that is capable of determining the expressivity and invertibility of multiple diff increments. The approach uses rules to transform atomic change operations into a smaller set of composite actions applicable to ontology sub-graphs. Finally, the authors exhibit the applicability of this method for an improved version management and annotation migration in collaborative ontology curation.

3. Challenges and open issues

Over the last ten years, research has covered most of the initial challenges that emerge in collaborative knowledge curation settings. We have witnessed: (i) increasingly versatile collaborative knowledge acquisition processes (especially in the context of Wikis and collaborative knowledge repositories); (ii) consistency-driven knowledge capturing workflows; and (iii) increasingly improved and intuitive user experiences. Nevertheless, we believe that we still need to address the biggest challenges in this domain, and hence, we are facing a series of critical open issues. Most of these open issues revolve around the temporal and evolution-related aspects of knowledge and include managing change, revision and inconsistent knowledge in collaborative environments, learning knowledge acquisition patterns from collaboration or discovering new knowledge from recurring inconsistencies. Evolving the biomedical annotations when the ontologies they depend upon have changed, plus understanding and quantifying the effect of these changes, are still open problems that scientists face in their work. We have also witnessed the creation of a variety of tools and software systems that are trying to address the same problem of collaborative knowledge curation, each having to confront the same challenges, and often coming to very similar solutions after significant effort and resources have been invested. A more principled and methodological approach that captures the lessons learned and synthesises the best practices into a series of well-defined steps would tremendously help the community by avoiding duplicative efforts and speeding the collaborative curation process. Finally, from an analytical perspective, we need to define and investigate metrics that measure the quality of the resulting curated knowledge, which will then enable an optimisation of the collaborative curation process.

References

[1] Bodenreider O, Stevens R. Bio-ontologies: current trends and future directions. Briefings Bioinformatics 2006;7:256–74.
[2] Staab S, Studer R, editors. Handbook on ontologies. Series on handbooks in information systems. Springer; 2009.
[3] Gibson F, Malone J. Community driven ontology development; 2010. <http://ontogenesis.knowledgeblog.org/217>.
[4] Consortium UniProt. The Universal Protein Resource (UniProt). Nucleic Acids Res 2005;33:154–9.
[5] Ashburner M, Ball C, Blake J, et al. Gene ontology: tool for the unification of biology. Nat Genet 2000;25:25–9.
[6] Fokkema I, Taschner P, Schaafsma G, Celli J, Laros JF, den Dunnen JT. LOVD v.2.0: the next generation in gene variant databases. Hum Mutat 2011;32:557–63.
[7] Croft D, OKelly G, Wu G, et al. pathways and biological processes. Nucleic Acids Res 2011;39:D691–7.
[8] Kanehisa M. Post-genome informatics. Oxford University Press; 2000.
[9] Hamosh A, Scott AF, Amberger JA, Bocchini C, Valle D, McKusick VA. Online mendelian inheritance in man: a knowledge base of human genes and genetic disorders. Nucleic Acids Res 2002;30:52–5.
[10] Pagon RA, Bird TD, Dolan CR, Stephens K, Adam MP. GeneReviews; 1993. <http://www.ncbi.nlm.nih.gov/pubmed/20301295>.
[11] Stokes TH, Torrance JT, Wang MD. ArrayWiki: an enabling technology for sharing public microarray data repositories and meta-analyses. BMC Bioinformatics 2008;9:S18.
[12] Bhardwaj A, Scaria V, Raghava GP, et al. Open source drug discovery – a new paradigm of collaborative research in tuberculosis drug development. Tuberculosis (Edinb) 2011;91:479–86.
[13] Vashisht R, Mondal AK, Jain A, Shah A, Vishnoi P, et al. Crowdsourcing a new paradigm for interactome driven drug target identification in Mycobacterium tuberculosis. PLoS One 2012;7:e39808.
[14] Heath T, Bizer C, editors. Linked data: evolving the web into a global data space. Synthesis lectures on the semantic web: theory and technology. Morgan & Claypool; 2009.
[15] Robinson PN, Kohler S, Bauer S, Seelow D, Horn D, Mundlos S. The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. Am J Hum Genet 2008;83:610–5.
[16] Smith B, Ashburner M, Rosse C, et al. The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 2007;25:1251–5.
[17] Gao Y, Kinoshita J, Wu E, et al. SWAN: a distributed knowledge infrastructure for Alzheimer disease research. Web Semantics Sci Serv Agents World Wide Web 2006;4:222–8.
[18] Ciccarese P, Wu E, Wong G, Ocana M, et al. The SWAN biomedical discourse ontology. J Biomed Informatics 2008;41:739–51.
[19] Groza T, Zankl A, Li Y-F, Hunter J. Using semantic web technologies to build a community-driven knowledge curation platform for the skeletal dysplasia domain. In Proc of the 10th International Semantic Web Conference (ISWC 2011). Bonn, Germany; 2011. p. 81–96.
[20] Groza T, Hunter J, Zankl A. The bone dysplasia ontology: integrating genotype and phenotype information in the skeletal dysplasia domain. BMC Bioinformatics 2012;13.
[21] Gruber TR. A translation approach to portable ontology specifications. Knowl Acquis 1993;5:199–220.
[22] Hoehndorf R, Bacher J, Backhaus M, Gregorio S, et al. BOWiki: an ontology-based Wiki for annotation of data and integration of knowledge in biology. BMC Bioinformatics 2009;10:S5.

[23] Schaffert S, IkeWiki: a semantic Wiki for collaborative knowledge management. In: Proc of the 1st international workshop on semantic technologies in collaborative applications. Los Alamitos, California; 2006. p. 388–96.
[24] Ghidini C, Rospocher M, Serafini L. Conceptual modeling in Wikis: a reference architecture and a tool. In: Proc of the 4th international conference on information, process, and knowledge management (eKNOW2012). Valencia, Spain; 2012.
[25] Hoehndorf R, Loebe F, Kelso J, Herre H. Representing default knowledge in biomedical ontologies: application to the integration of anatomy and phenotype ontologies. BMC Bioinformatics 2007;8.
[26] Schulz S, Beisswanger E, Wermter J, Hahn U. Towards an upper-level ontology for molecular biology. In: Proc of AMIA annual symposium. Washington, DC, USA; 2006. p. 694–98.
[27] Rospocher M, Eccher C, Ghidini C, Hasan R, Seyfang A, Ferro A, et al. Collaborative encoding of Asbru clinical protocols. In: Proc of eHealth 2010 – 3rd international ICST conference on electronic healthcare for the 21st century. Casablanca, Morocco; 2010.
[28] Tudorache T, Nyulas C, Noy NF, Musen MA. WebProtege: a collaborative ontology editor and knowledge acquisition tool for the Web. Semantic Web J 2012.
[29] Gennari JH, Musen MA, Fergerson RW, et al. The evolution of Protg: an environment for knowledge-based systems development. Int J Hum-Comput Studies 2003;58:89–123.
[30] Tudorache T, Falconer SM, Nyulas C, Noy NF, Musen MA. Will semantic web technologies work for the development of ICD-11? In Proc of the 9th International Semantic Web Conference (ISWC 2010) – In-Use Track. Shanghai, China; 2010. p. 257–72.
[31] Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas C, Tudorache T, et al. BioPortal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. Nucleic Acids Res 2011;39:W541–5.
[32] Malone J, Stevens R. Measuring the level of activity in community built bio-ontologies. J Biomed Informatics 2013;46:5–14.
[33] Hartung M, Gross A, Rahm E. COnto-Diff: generation of complex evolution mappings for life science ontologies. J Biomed Informatics 2013;46:15–32.