The ontology life cycle: Integrated tools for editing, publishing, peer review, and evolution of ontologies

Natalya Noy, PhD, Tania Tudorache, PhD, Csongor Nyulas, MS, Mark Musen, MD, PhD
Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, CA 94305

Abstract

Ontologies have become a critical component of many applications in biomedical informatics. However, the landscape of the ontology tools today is largely fragmented, with independent tools for ontology editing, publishing, and peer review: users develop an ontology in an ontology editor, such as Protégé; and publish it on a Web server or in an ontology library, such as BioPortal, in order to share it with the community; they use the tools provided by the library or mailing lists and bug trackers to collect feedback from users. In this paper, we present a set of tools that bring the ontology editing and publishing closer together, in an integrated platform for the entire ontology lifecycle. This integration streamlines the workflow for collaborative development and increases integration between the ontologies themselves through the reuse of terms.

Introduction

Ontologies in biomedicine facilitate information integration, data exchange, search of biomedical data, and other critical knowledge-intensive tasks.[1] Recent developments are dramatically changing the way that biomedical scientists are building terminologies and ontologies. As ontologies have become mainstream within biomedicine, they are being developed collaboratively by increasingly large groups of scientists. Furthermore, ontologies are becoming so large in their coverage (e.g., NCI Thesaurus with 80,000 concepts) that no single centralized group of people can develop them effectively and many organizations are inviting the broader user community to make contributions.

Even though specific collaborative workflows differ from one project to another, collaboration, discussion, and distributed contribution are present at many stages of the lifecycle of a biomedical ontology.[2] A number of tools today support collaboration at these different stages of the lifecycle. We give examples of such tools in the next section. However, these tools are generally disconnected from one another. Thus, when community members make contributions at one stage of the lifecycle (e.g., provide a comment on a published ontology), these contributions are not naturally integrated into the tools for another (e.g., ontology authors do not see these comments in their ontology editor).

In this paper, we discuss the tools based on the Protégé ontology-editing environment[3] and the BioPortal ontology library[4] that bring these stages of the ontology lifecycle "under one roof." We integrated Protégé and BioPortal in several ways to give users in a collaborative setting a more seamless experience. We have validated the integrated components by deploying them in several projects, including the development of the 11th Revision of the International Classification of Diseases (ICD-11) by the World Health Organization and the projects within the National Center for Biomedical Ontology (NCBO) and its collaborators.[5]

Tool support for the lifecycle of collaborative ontology development

The current landscape of tools that support various stages of collaborative ontology development includes tools for *ontology editing*, *ontology publishing,* and *collecting feedback from users.*

Ontology Editing Tools

Developers of biomedical ontologies use a number of tools to create and edit their ontologies. These tools include standalone tools such as OBO Edit,[6] which supports editing of ontologies in the OBO format. OBO Edit does not allow simultaneous editing by multiple authors, and developers usually use a source-code versioning system such as CVS to maintain and share different versions. The Protégé ontology editor, developed by our group, has several configurations, including Collaborative Protégé,[7] a desktop application that allows multiple users to edit an ontology simultaneously, discuss design decisions, make proposals, and analyze changes.

Ontology Publishing Platforms

Once ontology authors feel that they have a stable version of an ontology, they often need to publish it for other researchers to use or to provide feedback. *Ontology libraries* support ontology publishing by providing access to multiple ontologies and often enabling browsing, search, and other features. For example, obofoundry.org contains a collection of ontologies that are candidates to the OBO Foundry.[8]

The site uses CVS, a source-code versioning system, to store versions of the ontologies. The CVS infrastructure includes a bug tracker that many developers of OBO Foundry candidate ontologies use to track term suggestions and other issues with their ontologies. BioPortal, developed by NCBO, is a library where any biomedical ontology can be published. It currently hosts close to 200 biomedical ontologies. It enables users to browse ontologies, to search within and across them, to access records from biomedical resources that are annotated with ontology terms and to annotate their own experimental data with ontology terms.[9]

Collecting Feedback From Users

Developers of biomedical ontologies have a variety of ways to collect feedback from the user community once the ontology is published and others start to use it in their applications or to annotate their data. For example, many ontologies have a mailing list associated with them where the developers discuss design choices. Researchers also use the bug tracking mechanism in CVS to request new terms in ontologies, such as the Gene Ontology, and to track the progress of the request. Sometimes, developers simply share Microsoft Word documents or Excel spreadsheets, which contain, for example, definition of ontology classes. Collaborators make changes in these documents, using familiar formats, and a designated author then ports the changes to the ontology. This approach is occasionally used in the development of the Biomedical Resource Ontology.[10] LexWiki,[11] a tool based on Semantic MediaWiki, provides a mechanism for structured proposals for new terms or property values. These proposals are linked to the original classes that users propose to change. Developers then use an ontology editor, such as Protégé, to act on the requests.

These mechanisms for ontology editing, publishing and user feedback cover most of the functionality that users may need, but the tools are largely disconnected from one another. There is no easy way to publish an ontology from within an ontology editor, to find, extract, and include components of published ontologies in your own ontology, to see comments and requests from users in the context of a class that you are editing.

WebProtégé and BioPortal

In this paper, we focus on two tools developed in our laboratory, WebProtégé and BioPortal, and the integration between them. The tools support ontology editing, publishing, and collection of user feedback.

WebProtégé[12] is a Web-based version of the Protégé ontology editor. It provides a simple user interface to edit an ontology using a Web browser, enables multiple users to access and edit the same ontology simultaneously, and allows users to configure their own interface. Users can add threaded discussions to any ontology element to discuss their design or to record provenance of a class. WebProtégé also provides a view of change history.

BioPortal is an open library of biomedical ontologies. BioPortal, uses the social approaches in the Web 2.0 style to bring structure and order to the collection of biomedical ontologies. It enables users to provide and discuss a wide array of knowledge components, from submitting the ontologies themselves, to commenting on and discussing classes in the ontologies, to reviewing ontologies in the context of their own ontology-based projects, to creating mappings between overlapping ontologies and discussing and critiquing the mappings.[13]

The two tools support the different stages of the ontology lifecycle (Figure 1): Ontology authors use Protégé to develop their ontology collaboratively, with discussions among themselves. They use BioPortal to publish the ontology and to solicit feedback from the broader user community. Both tools provide Application Programmer Interfaces (APIs) to enable the use of ontologies in other applications: Protégé provides a Java API and BioPortal provides a REST Web service API.

A use case scenario

We will now describe a typical use case that the components described in this paper would support:

1. A group of developers creates the first version of their ontology in Protégé. During development, they include references to terms in other ontologies.

2. The authors publish their ontology in BioPortal. At the same time, they create a copy of the ontology. This new copy will be the current working copy where the authors will continue to make edits while others access a stable version through BioPortal.

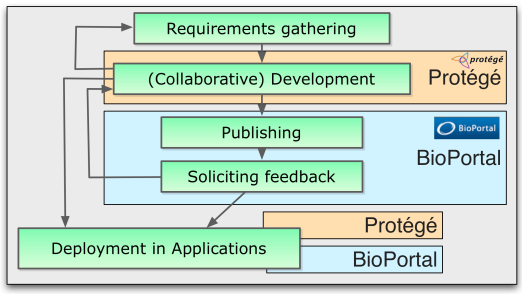

Figure 1. Protégé and BioPortal in the ontology lifecycle. Protégé supports ontology development; BioPortal supports ontology publishing and collection of feedback from the users.

3. The authors send the link to the newly published ontology to the user community, asking for feedback.

4. The community members use the link to go directly to the ontology in BioPortal. They may look at the metadata that the authors provided, metrics that BioPortal computed, and navigate around the tree hierarchy or neighborhood view.

5. Users add notes to BioPortal, commenting on specific classes and requesting new terms.

6. Ontology authors get notified through an RSS feed that there are new comments on their ontology. The feed contains a link to the specific concepts that had comments. They edit the ontology in Protégé, based on the comments.

7. The authors publish a new version in BioPortal. Users who have subscribed to the RSS feed for this ontology get notified that a new version is available.

Bringing publishing and editing together

We developed a set of tools that bridge the different elements of the ontology lifecycle, enabling most steps in the scenario above. While we do not have a completely integrated platform yet, these tools bring us very close to an integrated set of tools.

Using published ontologies to populate your ontology

Ontology reuse is critical if we are to achieve the true potential of ontologies as enablers of data integration. For example, if SNOMED CT defines a hierarchy of diseases, or the Foundational Model of Anatomy defines the vocabulary for human anatomy, other ontologies that need these vocabularies need to be able to reuse them easily rather than develop their own. The **BioPortal reference widget** (Figure 2) enables WebProtégé users to search ontologies in BioPortal and to include references to the terms from other ontologies. Protégé stores the reference as an object in the user ontology, which encapsulates the source of the term and the provenance information on the inclusion. The widget can be configured to search all of BioPortal or only a specific ontology. For instance, if the widget fills in the value for *Body Part* in ICD-11, we can configure it to search only the anatomy branch of SNOMED CT.

Users can designate any subset of an ontology that is published in BioPortal as a set of allowed values for the reference widget (a value set). We refer to these subsets as **ontology views**.[14] Anyone can contribute a view. For instance, for the *Body Part* property in Figure 2, we created a subset of SNOMED CT that included only the anatomy branch and configured the widget to search only this view.

Publishing ontologies in BioPortal from Protégé

After an ontology developer has created a new version of the ontology in Protégé, she can publish it directly in BioPortal. This functionality is currently in a prototype form. It uses the BioPortal Web service interface to add an ontology and its metadata to the BioPortal collection.

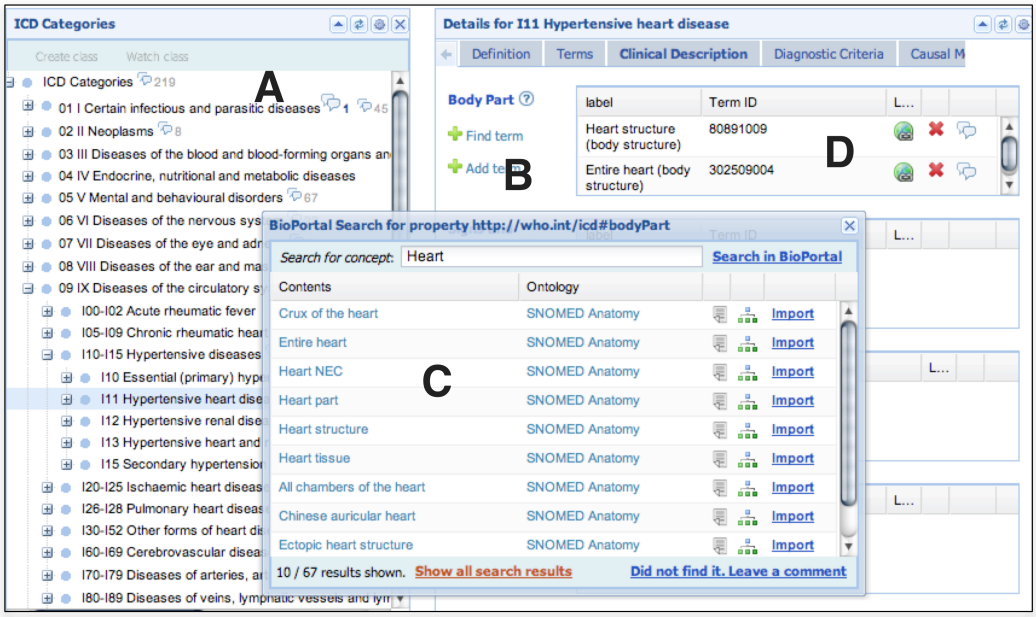

Figure 2. BioPortal reference widget. A user selects a class in the hierarchy (A). In order to fill in the value for *Body Part* (B), she uses the widget to search BioPortal (C) and to select the appropriate body part from SNOMED CT. The reference is then added as a link to SNOMED CT.

Sharing user feedback between BioPortal and Protégé: Structured Notes

Both BioPortal and Protégé enable their users to comment on classes in ontologies, by creating notes. The Protégé users often use notes to discuss the design of a class and to reach consensus on the decisions.[7] BioPortal users use the notes to provide feedback to the ontology developers on missing terms and other problems with the ontology.[13] We are integrating the two notes mechanisms to enable editors—in the context of their ontology-editing session—to see easily the notes that the broader user community has provided. Specifically, we have implemented a **Notes API**—a module that both BioPortal and Protégé will share. The Notes API encapsulates the representation, storage, and access mechanism for creating notes linked to ontology elements. The Protégé server that uses the same instance of the Notes API as BioPortal will provide its users a view of BioPortal notes directly in Protégé. Thus, when a user makes a term request in BioPortal, the ontology author will see the term request when editing the ontology in Protégé. After acting on the request, the author can change the status of the request as "completed"; or he can "archive" a request if it is no longer relevant. The BioPortal code will use the status to determine whether to display the note to its users. We have analyzed the requirements for notes in BioPortal and Protégé to develop an ontology of notes, including different types of proposals, such as term requests, property value changes, hierarchy changes, and retirement of concepts. The Notes API stores the notes as instances of these ontology classes.

Common implementation infrastructure

In addition to giving Protégé users direct access to library tools and enabling Protégé and BioPortal to share the notes, the tools also benefit greatly by the two types of application sharing other elements of common infrastructure.

Using Protégé and LexGrid to browse and display ontologies in BioPortal

BioPortal is different from many ontology libraries in that it not only lists the ontologies, but also provides search and browsing of the ontologies. In order to support ontology browsing, BioPortal needs to have fine-grained access to ontology components—the access usually provided by ontology-editing tools. BioPortal uses two ontology-editing APIs to get detailed information about ontology elements: it uses the Protégé API for OWL and Protégé frames ontologies and the LexGrid API for the ontologies in RRF and OBO formats.[15] BioPortal also uses these APIs to get additional information about its ontologies, including computing metrics, such as the number of classes and properties, and level of conformance to best practicecs.

Using ontologies as infrastructure for BioPortal

Using ontologies to represent application data and to drive knowledge infrastructure of software projects provides separation of the declarative and procedural knowledge and allows independent evolution of the declarative knowledge. BioPortal uses the BioPortal Metadata Ontology to represent details about all the ontologies in the repository, internal system information and the user generated information such as mappings between classes in different ontologies and ontology reviews. This representation "drives" the application and BioPortal is the first large-scale application that uses ontologies to represent essentially all of its internal infrastructure.[14]

Technology

BioPortal provides programmatic access to all its content through REST Web service APIs.[9] The BioPortal reference widget uses the Search web service to search the ontologies. The Search web service can take a parameter limiting the search to a specific ontology or a set of ontologies. The widget then uses the term services to access details of the term that the user has selected. It uses a FlexViz plugin, developed at the University of Victoria to display the neighborhood of the selected term. The facility to publish ontology from Protégé to BioPortal uses the POST Web service to post a new ontology.

Protégé and LexGrid provide Java APIs that BioPortal uses to parse ontologies and to access the details about them when users navigate the site. We have implemented high-level Java APIs to create, access, and update notes and to access metrics.

Validation

The tools that we described in this paper are used in production environment, thus providing validation of the approach. BioPortal gets 11,000 visitors each month, with 30,000 page views. WebProtégé is used as the primary editing environment in several large-scale projects. Specifically, iCAT, a custom-tailored version of WebProtégé configured with a large number of BioPortal reference widgets, is currently the primary editing environment for ICD-11 at WHO. It currently has 65 registered users, who have made over 2100 changes and added over 1150 notes.

Discussion

We have described a simple editing and publishing workflow in this paper. In practice, these workflows

are often more complex and they differ from one project to another. We are working on supporting elements or a more flexible and customizable workflows. We currently have some elements of this support, including fine-grained access policies for users in Protégé and task lists. Workflows that integrate both editing and publishing will necessarily be more complex.

Protégé and BioPortal have different sets of user accounts. However, in order for the integration to be truly seamless, users must be able to preserve their identity from one tool to another. We are currently implementing the use of OpenID protocol for both tools, which would enable users to login to both Protégé and BioPortal with the same credentials.

In most workflows, there will be at least two "active" versions of an ontology: one is a published version that users can view and comment on and another is the "working" version where authors are applying their changes. Thus, our mechanism for attaching notes to classes must be flexible enough to enable notes to show up in both cases. Notes must have both a reference to a class in a specific version (e.g., the published version, which was the one the user saw when she created a request), and a global ontology id, enabling the editing tool to access this note, even when the users are working with a different version. We use this approach in our Notes API.

Because both Protégé and BioPortal have open APIs, several other groups have integrated elements of ontology editing and publishing using these APIs. For instance, there are Protégé plugins that access other ontology libraries, such as TONES[16] and Watson[17] (neither of which is specific to biomedicine), and enable users to include ontologies or terms from other libraries. The OntoFox application[18] enables users to extract terms and their description from a pre-defined set of 15 ontologies and save the extracted information in a file, based on the MIREOT guidelines.[19] A user ontology can then import this file to include the referenced terms.

Conclusion and future plans

We have described a number of components that bridge the gaps in the ontology editing and publishing lifecycle that enable better integration of the various stages of collaborative distributed community-based ontology development. Our future work includes implementation of the remaining components of this integrated platform, including direct access to editing features from BioPortal and the use of the Notes API in both tools.

Our experience with the development of ICD-11 and our interactions with many NCBO collaborators, including such projects as NIF,[20] NEMO,[21] shows that users today require and benefit from integration of editing and publishing. These projects drive our requirements for the tools for the integrated lifecycle.

References

1. Rubin, D.L., Shah, N.H., and Noy, N.F., Biomedical ontologies: a functional perspective. Brief Bioinform, 2008. 9(1): p. 75-90.
2. Sebastian, A., et al. A Generic Ontology For Collaborative Ontology-Development Workflows. 16th Intl. Conf. on Knowledge Engineering (EKAW). 2008. Catania, Italy: Springer.
3. Protégé. http://protege.stanford.edu. Accessed 6/2010.
4. BioPortal. http://bioportal.bioontology.org. Accessed 6/10.
5. The National Center for Biomedical Ontology (NCBO). http://www.bioontology.org/. Accessed 6/10.
6. OBO-Edit http://oboedit.org/. Accessed 6/2010.
7. Noy, N.F., et al. Developing Biomedical Ontologies Collaboratively. AMIA 2008 Annual Symposium. 2008. Washington, DC.
8. Smith, B., et al., The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotech, 2007. 25(11): p. 1251-5.
9. Noy, N.F., et al., BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res, 2009. 37(Web Server issue): p. W170-3.
10. Biomedical Resource Ontology. http://purl.bioontology.org/ontology/BRO.
11. LexWiki. https://cabig-kc.nci.nih.gov/Vocab/KC/index.php.
12. Tudorache, T., Vendetti, J., and Noy, N. WebProtégé: A Lightweight OWL Ontology Editor for the Web. 5th Int. Workshop on OWL: Experiences and Directions (OWLED). 2008. Karlsruhe, Germany.
13. Noy, N.F., et al. Harnessing the Power of the Community in a Library of Biomedical Ontologies. Workshop on Semantic Web Applications in Scientific Discourse at ISWC. 2009. Chantilly, VA.
14. Nyulas, C.I., et al. Ontology-Driven Software: What We Learned From Using Ontologies As Infrastructure For Software. Workshop on Semantic Web Enabled Software Engineering at ISWC. 2009. Chantilly, VA.
15. Pathak, J., et al., LexGrid: A Framework for Representing, Storing, and Querying Biomedical Terminologies from Simple to Sublime. J. American Med. Informatics Assoc, 2009. 16(3): p. 305--315.
16. TONES. http://owl.cs.manchester.ac.uk/repository/.
17. d'Aquin, M., et al. Watson: A Gateway for Next Generation Semantic Web Applications. Poster session at ISWC 2007. Busan, Korea.
18. OntoFox. http://ontofox.hegroup.org. Accessed 6/10.
19. Courtot, M., et al. MIREOT: the Minimum Information to Reference an External Ontology Term. Int. Conf. on Biomed. Ontology (ICBO) 2009.
20. Neuroscience Information Network (NIF). http://www.neuinfo.org/. Accessed 6/10.
21. Neural ElectroMagnetic Ontologies (NEMO). http://nemo.nic.uoregon.edu/wiki/NEMO. Accessed 6/10