

Protégé 5.5 – Improvements for Editing Biomedical Ontologies

Matthew Horridge, Rafael S. Gonçalves, Csongor Nyulas,
Tania Tudorache and Mark A. Musen

Stanford Center for Biomedical Informatics Research, Stanford, California, USA

1 INTRODUCTION

We present Protégé 5.5, a significant update to the Protégé Desktop software, which contains new features that are geared towards editing biomedical ontologies. This version of Protégé contains user-interface enhancements and optimizations that should make the browsing and editing of OBO-library-style biomedical ontologies easier, faster and more efficient when compared to previous versions of Protégé.

2 NEW FEATURES

In what follows we describe the major features that are included with this release of Protégé. Inspiration for these features came directly from the biomedical ontology community.

2.1 Viewing and Searching for Terms

This version of Protégé contains various enhancements related to the display of terms in various views:

Display of OBO Ids Protégé now displays OBO-style identifiers alongside human readable labels in various places throughout the user interface. This includes presentation in the search results view (Fig. 1), in tooltips that are shown when hovering over terms (Fig. 2) and also in the term browsing view (Fig. 3).

Links in Annotation Values We have expanded the kinds of tokens in annotation and XRef values that get rendered as clickable links. In this latest version of Protégé, substrings that correspond to PubMed Ids (e.g. “PMID:18461074”), Wikipedia pages (e.g. “Wikipedia:DNA_repair”) and ISBNs (e.g. “ISBN:0198506732”) are clickable and link out to the appropriate Web pages. In the future, we intend to offer some form of customization of this functionality so that other database identifiers can be made clickable.

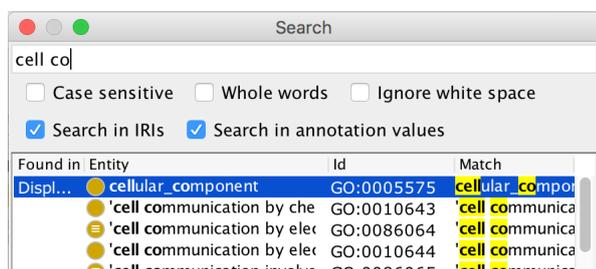


Fig. 1: The Search Results View now displays OBO Ids for terms that have IRIs corresponding to OBO Ids. Multi-word search is also now supported.

Search Improvements The default search functionality in Protégé now supports multi-word search that is insensitive to the order in

which search words are entered. For example, searching for “Heart Valve” will now return terms whose labels contain “Heart Valve” and also “Valve of Heart”. Search can also be performed on OBO Ids without changing any configuration settings (e.g. the renderer). For example, searching the Gene Ontology (GO) for “GO:0005634” will find the term nucleus.

2.2 Browsing Hierarchical Structure

Biomedical ontologies tend to be large, with hierarchies that are deep and that have large spans. Furthermore, hierarchies are multi-axial, with relationships other than IS-A (SubClassOf) being important. We have made two major improvements with regards to hierarchy browsing:

Mixed Hierarchy Viewing We have added the ability to display relationships other than IS-A in the class hierarchy view. This long-awaited feature can be toggled on or off via the View menu. Figure 2 shows an example mixed hierarchy from the GO. SubClass child-parent relationships are distinguished from non-subclass parent-child relationships using differently colored arrows — blue arrows represent non-subclass relationships.

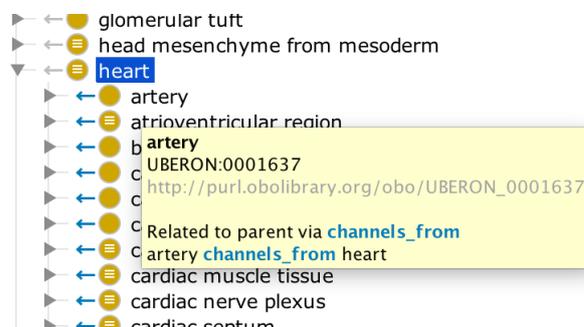


Fig. 2: An example of mixed relationship hierarchies being displayed in the class tree. Blue arrows indicate a child-parent relationship other than IS-A (SubClassOf). Tooltips display more information about how a child is related to its parent—here the mouse is hovering over artery.

Hovering over a class in the hierarchy displays a tooltip stating the relationship that causes the child-parent relationship to hold. In the future we intend to allow subsets of properties to be used in the display of the tree. We are also considering using customizable colors in order to provide more of a visual clue as to the underlying relationship type between child-parent terms.

Breadcrumb Trails When browsing large hierarchies it is easy to lose sight of the positions of selected terms and the branch that they appear in. We have therefore added a breadcrumb trail to the top of the main Protégé window. This can be toggled on or off as necessary.

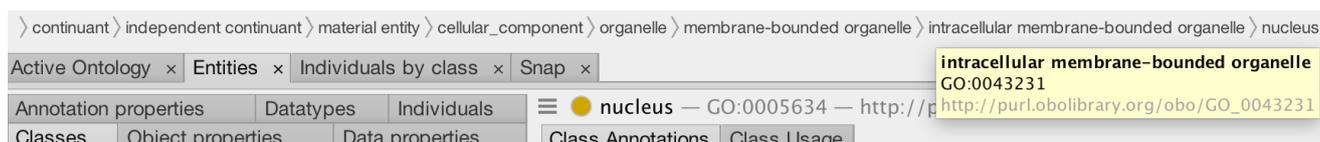


Fig. 3: Breadcrumb and Term Header. A clickable breadcrumb trail is displayed above all tabs that shows the path the currently selected term. OBO Ids are now displayed beside the label/IRI for the currently selected term. The three bar button, next to the term icon, exposes a menu containing items that can be used to copy OBO Ids, copy the term as markdown (for use in GitHub issues for example) and navigate to the term IRI in a Web browser.

Hovering over a term name in the breadcrumb trail shows the display name for the term along with its OBO Id and its IRI. Term names can be clicked to change the term selection in order to jump up to parent and ancestor terms in the hierarchy.

2.3 Workflow Support for Obsolete Terms

One significant new feature that we have added is a wizard for obsoleting (deprecating) terms. It is typically the case that several detailed steps are involved in the deprecation of terms. Furthermore, each ontology project can have slightly different deprecation strategies. We have therefore provided a configurable deprecation wizard that leads users through the steps that are required when deprecating terms. Deprecation profiles, which are simple YAML files, can be dropped into a Protégé installation in order to provide project specific deprecation workflows. We have provided three profiles out of the box (1) A generic “Basic” profile; (2) A “Gene Ontology (GO)” profile, that follows the steps required by the GO consortium¹; and (3) A “Ontology for Biomedical Investigation (OBI)” profile, that is as detailed as the GO profile, but requires different steps.

2.4 Integration with Git

Biomedical ontology projects have long used issue trackers and version control systems for project management. In recent years GitHub has come to dominate the version control repository landscape and projects such as the GO project now use this technology. The GO workflow encourages editors to create Git branches for additions and bug fixes. These branches are then merged into the master branch once the editing for a particular feature is complete. We have therefore provided a lightweight integration of Protégé with Git.



Fig. 4: Git repository status as displayed in the lower left corner of the Protégé window. The current branch (master) is shown along with the status of ontologies in the imports closure of the loaded ontology.

Display of Git Repository Status If Protégé detects that an ontology has been loaded from a location within a Git repository, it will display the checked out branch on the left hand side of the status

¹ <http://go-ontology.readthedocs.io/en/latest/ObsoleteTerm.html>

bar (Fig. 4). It will also display an indication as to whether there have been uncommitted changes to the ontologies in the imports closure of the loaded ontology.

Use of Git User Name for Term Creation Metadata The last few versions of Protégé have offered the ability for various term creation metadata to be automatically stamped onto a term as annotations. One auto-generated annotation is “created by”, which takes the name of an editor. In this latest version, Protégé can now utilize the Git user name for the current Git repository (Fig. 5), so that “created by” annotation values correspond to Git user names in commits to a repository.



Fig. 5: User name preferences. Protégé now allows the user name to be derived from the active Git repository so that term creation metadata is aligned with the user name in repository commits.

3 CONCLUSIONS

We have presented the major features of Protégé 5.5 that are geared towards the editing of biomedical ontologies. We seek feedback from the biomedical ontology community in order that we may continue to enhance Protégé so that it meets the needs of this community.

ACKNOWLEDGEMENTS

We are heavily indebted to the Biomedical Ontology Community for suggestions and feedback on the new features that we have presented in this paper. We would especially like to thank the participants of the 2017 Berkeley GO Editors Workshop: (in alphabetical order) Kimberly Van Auken, Tanya Berardini, Seth Carbon, Karen Christie, Eric Douglass, Pascale Gaudet, Harold Drabkin, David Hill, Huaiyu Mi, Suzi Lewis, Donghui Li, Chris Mungall, Monica Munoz-Torres, David Osumi-Sutherland, and Nicole Vasilevsky – this work would not have been possible without their input. Special thanks go to Chris Mungall for an invitation to this meeting.

This work was funded by Grant GM121724 from the National Institute of General Medical Sciences at the United States National Institutes of Health.