Bioinformatics, 33(1), 2017, 148–149
doi: 10.1093/bioinformatics/btw579
Advance Access Publication Date: 6 September 2016
Applications Note

Databases and ontologies

OntoBrowser: a collaborative tool for curation of ontologies by subject matter experts

Carlo Ravagli, Francois Pognan and Philippe Marc*

PreClinical Safety, Translational Sciences, Novartis Institute for Biomedical Research, Basel, CH-4002, Switzerland

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso

Received on November 5, 2015; revised on July 8, 2016; accepted on September 2, 2016

Abstract

Summary: The lack of controlled terminology and ontology usage leads to incomplete search results and poor interoperability between databases. One of the major underlying challenges of data integration is curating data to adhere to controlled terminologies and/or ontologies. Finding subject matter experts with the time and skills required to perform data curation is often problematic. In addition, existing tools are not designed for continuous data integration and collaborative curation. This results in time-consuming curation workflows that often become unsustainable. The primary objective of OntoBrowser is to provide an easy-to-use online collaborative solution for subject matter experts to map reported terms to preferred ontology (or code list) terms and facilitate ontology evolution. Additional features include web service access to data, visualization of ontologies in hierarchical/graph format and a peer review/approval workflow with alerting.

Availability and implementation: The source code is freely available under the Apache v2.0 license. Source code and installation instructions are available at http://opensource.nibr.com. This software is designed to run on a Java EE application server and store data in a relational database.

Contact: philippe.marc@novartis.com

1 Introduction

Many code lists and ontologies have been created to model biological concepts. Databases are able to consolidate and integrate data from multiple sources by adhering to controlled terminologies and ontologies. Contributors to such databases are generally required to submit data according to a compatible vocabulary (Côté *et al.*, 2010; de Coronado *et al.*, 2004; Smedley *et al.*, 2015). However, biological results are often captured using inconsistent nomenclatures and/or vocabularies incompatible with the target databases. In order to achieve data consistency and compatibility, nomenclature from the original data must be mapped to a target ontology or code list. This translation task needs to be conducted by domain experts. Typically ontology creation tools like WebProtégé (Horridge *et al.*, 2014) are not designed for this type of task or user. Other tools such as Karma (Szekely *et al.*, 2014) are designed specifically for mapping/alignment and address the initial integration problem (see here for a review of useful tools: http://www.mkbergman.com/1769/50-ontology-mapping-and-alignment-tools/). However, none of the tools are designed to be part of an ecosystem facilitating continuous data integration. Consequently, the vocabulary mapping/translation task and consequent ontology evolution are often performed using snapshots of exported data followed by reconciliation. Continuous data integration coupled with evolving ontologies was a major challenge faced by the Innovative Medicines Initiative eTOX consortium (Cases *et al.*, 2014). Results from over 6000 toxicology reports were manually extracted over a 5-year period. The original reports, generated over several decades, were contributed by 13 independent pharmaceutical companies and hence written using many different nomenclatures. The complexity and scale of the challenge was addressed by developing the OntoBrowser tool. The tool has matured over 4 years and has been collaboratively used by over a dozen consortium domain experts to map more than 70 000 distinct terms to 6352 preferred ontology or code list terms.

2 Using OntoBrowser

2.1 Online collaborative curation

It is common for multiple curators, potentially located at different sites, to work collaboratively. The majority of tools for ontology

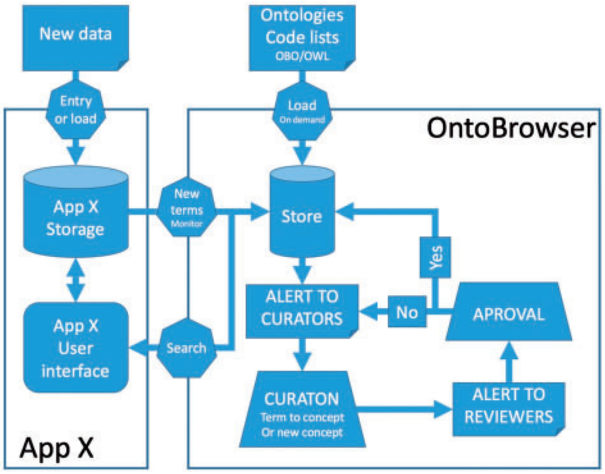

Fig. 1. OntoBrowser integration in an ecosystem. OntoBrowser internal mechanics (right), supplying web services to another application (left)

curation and/or mapping of reported terms to controlled terminologies are deployed locally and restricted to manipulating data in isolation. This leads to multiple local copies of the data being modified independently and hence requires merging at each milestone. The peer-review process also requires careful coordination. Even with the correct file formats, tools, and procedures in place, reconciliation and coordination can be time-consuming and error prone. Furthermore, it adds additional unnecessary effort/overhead and complexity to the curation process. OntoBrowser was specifically designed for multi-user online collaboration and peer review. A central database hosts all the data (i.e. multiple ontologies and code lists) providing a single working copy shared by all users. As a web-based application, it can be deployed on a server accessible via the internet (like the eTOX instance) or within an intranet.

The user interface has been developed in close collaboration with multiple biologists to ensure that the design is both logical and efficient. It allows searching and browsing of the concepts. The user interface supports a read-only mode and a curation mode, depending on the privileges defined for the user. The curation mode exposes functionality for modifying ontologies, mapping report terms and approving (or rejecting) pending changes. The peer review workflow is implemented as part of the core application functionality (Fig. 1). E-mail alerts are sent to curators when pending changes are outstanding and require approval. Other features include versioning and a complete audit history.

Another key feature of the software is the automatic pre-mapping of unmapped reported terms to ontologies (or code lists). The logic includes stemming and ignores the order of words to provide fuzzy matching. The automated fuzzy matching pre-mapping greatly reduced the curation work required by the scientists during the eTOX project.

2.2 Web services enabling system integration

OntoBrowser provides web services to expose ontology data and application functionality to other applications or services. For example, Novartis utilized OntoBrowser web services in its Translational Safety Platform (TSP) data warehouse to develop an interactive histo-pathology search application enabling users to query microscopic findings using multiple ontology terms. These findings are continually consolidated from multiple source systems into a data warehouse. OntoBrowser is used by domain experts to map the tissue and histopathology vocabularies to two respective ontologies. At runtime, the TSP frontend calls OntoBrowser's *search* and *ontology visualisation* web services to provide a user interface to create search criterion, allowing users to search and browse the anatomy and histopathology ontologies directly within the TSP application (Fig. 1). The TSP backend calls OntoBrowser web services to retrieve a list of subclasses of the ontology terms selected by the user to query the data warehouse.

Using the web services, ontologies (optionally including synonyms) can also be fully exported from OntoBrowser. Several standard ontology formats are supported to ensure interoperability with other tools/systems, e.g. OWL (RDF and XML), OBO, Manchester and Turtle.

3 Installing OntoBrowser

OntoBrowser requires a Java EE application server (e.g. Wildfly or WebLogic) and a relational database (e.g. MySQL or Oracle). Setting up a new instance, including the initial load of ontologies and connection, takes approximately 2 h. A full installation guide is provided in the source code repository. Ontologies from the public domain as provided by the OBO Foundry (Smith *et al.*, 2007) can be easily imported and synchronized.

Acknowledgements

We would like to acknowledge the contribution of many scientists from the eTOX project, especially Alessandro Piaia, Anne Ryan, Manuela Stolte, Esther Sutter, Peter Clemens and David Heard.

Funding

This work was supported by the Innovative Medicines Initiative Joint Undertaking [115002], resources of which are composed of financial contribution from the European Union's Seventh Framework Programme [FP7/2007-2013] and EFPIA companies' in kind contribution.

Conflict of Interest: none declared.

References

Cases,M. *et al.* (2014) The eTOX data-sharing project to advance in silico drug-induced toxicity prediction. *Int. J. Mol. Sci.*, **15**, 21136–21154.

Côté,R. *et al.* (2010) The Ontology Lookup Service: bigger and better. *Nucleic Acids Res.*, **38**, W155–W160.

de Coronado,S. *et al.* (2004) NCI Thesaurus: using science-based terminology to integrate cancer research results. *Stud. Health Technol. Inform.*, **107(Pt 1)**, 33–37.

Horridge,M. *et al.* (2014) WebProtégé: a collaborative Web-based platform for editing biomedical ontologies. *Bioinformatics*, **30**, 2384–2385.

Szekely,P. *et al.* (2014) Publishing the data of the Smithsonian American Art museum to the linked data cloud. *IJHAC*, **8**, 152–166.

Smedley,D. *et al.* (2015) The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res.* **43**(W1), W589–W598.

Smith,B. *et al.* (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.*, **25**, 1251–1255.