COBrA-CT: An e-Science tool for ontology curation

BBSRC e-Science development grant BB/D006473/1

Bonnie Webber, Jonathan Bard, Wenfei Fan, Stuart Aitken

Informatics, University of Edinburgh

The problems we address are version management and quality control in bio-ontologies that have an active contributing community, such as the many ontologies published the Open Biological Ontologies initiative. The version and quality control methods we propose may be relevant to the Gene Ontology as well.

The number of bio-ontologies is growing rapidly, indicating their important role in bioinformatics. But lacking is support for their curation – in particular for version management and quality control. Tools for ontology management are required in order to deal with the continuous process of ontology revision – no bio-ontology that is in active use remains unchanged after publication. MeSH is revised annually with a major sweep of MedLine, while the Gene Ontology (GO) is augmented and changed at much shorter intervals. Since biological data is annotated with respect to the terms (concepts) from a particular version of an ontology, updating data annotation to reflect these changes and additions requires effective management of different versions of an ontology, efficient archiving of previous versions, and the ability to wind forward and back among these versions. Quality control of ontologies involves, inter alia, error checking of proposed changes, and communication of the rationale for the proposed changes.

Our objective is to develop tools and techniques that will support both the curators of ontologies, and the community of biologists who are suggesting changes and additions to the ontology. This will be achieved through a server-based infrastructure that supports archiving and version control, along with an ontology editing tool with built-in error checking that will also allow users to provide the rationale, in the form of meta-data, for proposed ontology edits. Curators will have access to this meta-data along with the proposed changes. To further support the curator, the tool will provide a graphical visualisation of ontology changes. The ontology tool will be based on, COBrA (Aitken et al., 2005a), developed by the BBSRC-supported XSPAN project. Ontology version control and archiving will be done using the XML key techniques of Buneman et al. (2002).

At present, bio-ontology editors provide only the simplest version control (based on CVS) and no tool support is available for the conceptual analysis of ontologies. The adoption of Description Logics (DL) will permit efficient logical consistency checking and classification, but many of the Open Biological Ontologies (OBO) are far from this level of formalisation. Our proposals address the intermediate task of identifying and correcting oversights such as failing to provide a textual definition of a concept, and simple logical errors such as cycles in the class hierarchy, and contradictions arising from disjointness assertions.

The COBrA Curation Tool, COBrA-CT, will be developed in a modular fashion, meaning that modules will be capable of operating as Protégé plug-ins, as well as modules in the original COBrA framework. Making the new functionality available to Protégé users will widen the potential uptake of the results and achieve compatibility with existing e-Science ontology initiatives, for example, the Collaborative Open Ontology Development Environment project (CO-ODE) that has adopted this mode of delivery.

COBrA-CT (and a COBrA-enabled Protégé) will:

- Allow users to create, edit and explore an ontology, and supply rationale for edits that will be proposed to the curators (i.e. meta-data in the form of mappings between ontology versions);
- Allow users to create and review mappings between two ontologies (which may be successive version of the same ontology, or comparable ontologies such as anatomy ontologies of different organisms;
- Assist users to review ontological modelling decisions;

- Interact with an ontology management server to archive and retrieve ontologies, thus allowing multiple users and curators to coordinate their activities;

The ontology management server will:

- Organise version control and archiving using an XML key-based technique;
- Implement an authority model and a process model to organise curation and publication according to best practice.

Programme of Research

Curation has been recognised as a priority for e-Science (Lord and Macdonald, 2003), and is an important concern for many communities and in standards initiatives. For example, there are efforts to standardise the names used for tissue samples assayed by microarray (Parkinson, 2004), as well as the metadata that describes the experimental results (MGED/MIAME[1]). Ontologies are of central importance in curation, as only by defining the meaning of the terms used to describe a particular field can the underlying concepts be clarified and agreed upon among the research community, and used consistently for annotating data. A consistent, shared ontology is of critical importance to the sharing of knowledge, and has long-term value in supporting a systems-level approach to biology. For example, the Gene Ontology is in widespread use for data mining and data visualisation, and has great potential for further integration of data across the different levels of biological granularity. However, ontologies are not static: they must change to reflect changes in science, to adapt to new uses, to broaden their community or to remedy flaws. Ontologies have also been identified as key resources in numerous e-Science projects, including AstroGrid[2], MyGrid[3] and the Advanced Knowledge Technologies IRC[4].

We view the curation of ontologies in the context of e-Science as encompassing creating and publishing ontologies, as well as tracking changes and maintaining consistency in the ontologies after publication. In addition to version management, curation also includes the review of the content of the ontology, and assessment of quality. A related issue is the maintenance of ontological annotations assigned to data under a given ontology, as the ontology may change after a term has been used as an annotation and therefore one may wish for the annotation to be updated as well.

As the use of ontologies widens, the problems of tracking versions, and the changes between versions, and of reconciling differences in conceputal modelling arise. Addressing these are our main goals. We propose a server-based model for curation that allows remote users to create and submit annotated changes to ontologies and also to participate in the review process by applying some simple critiquing techniques that help identify errors. The proposed COBrA-CT will also support the curator by providing the appropriate management support and visualisations.

Approach

The current version of COBrA is an editor and mapping tool for ontologies in GO and OBO formats (see Figure 1). COBrA also allows users to explore two ontologies simultaneously, to make links between them and annotate those links with respect to a third ontology, as shown in Figure 2. COBrA was designed as a knowledge acquisition tool, to be used manually, in the context of making links between the anatomy ontologies of different model organisms. Throughout its development we have been aware that we are addressing specific instances of more general problems, namely ontology editing, mapping and management. This proposal is to extend and enhance COBrA by adding a range of ontology curation functions, and by enabling GRID and Web Service compatible modes of operation. The key element is to decentralise ontology curation (meaning curation in the wider sense of a community activity), while maintaining the integrity of the centrally-held ontology and mapping resources. COBrA-CT will operate as a stand-alone desktop tool, as at present, but will also be capable of interacting with an ontology management server which will be hosted by the Edinburgh Centre for

[1] www.mged.org

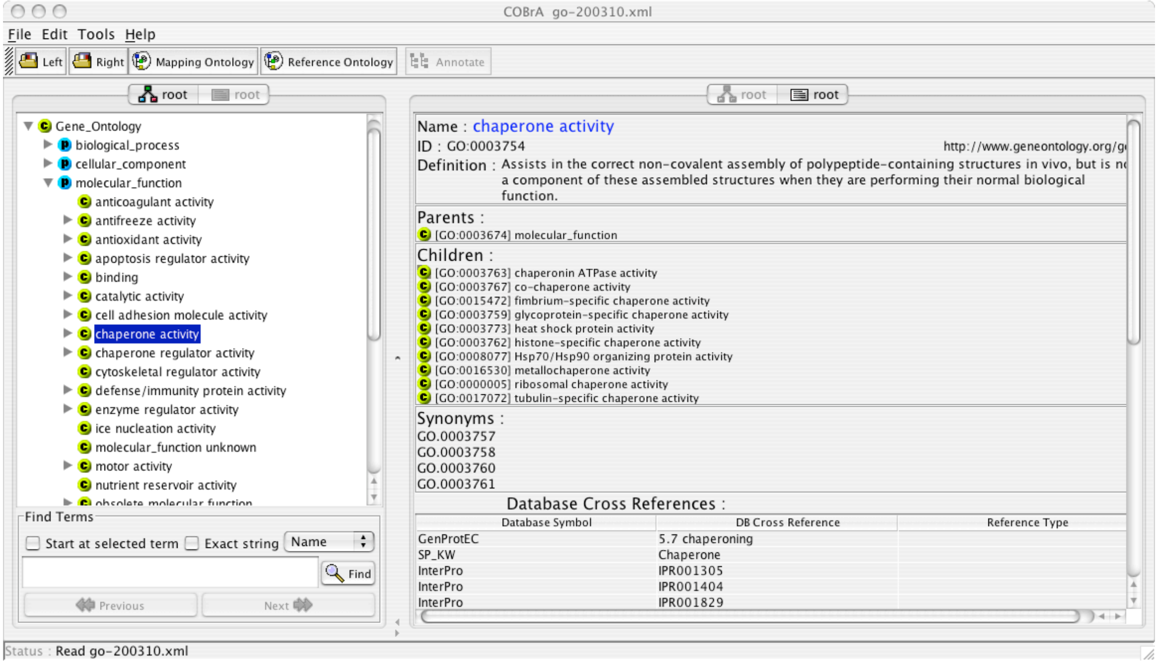
[2] astrogrid.semanticweb.org

[3] www.mygrid.org.uk

[4] www.aktors.org

Bioinformatics. The server will act as an archive, and will support an authority model by which curators and users have varying levels of access and publication rights.

The code to be developed will be designed as modules that will be built-in to COBrA, and will have an API to allow them to function as Protégé plug-ins. This will yield compatibility with other e-Science work that utilises Protégé - specifically that which is concentrating on OWL. Although Protégé already has a user base in the life sciences, its use does not appear to be widespread in the OBO community.



Name : chaperone activity
ID : GO:0003754 http://www.geneontology.org/g
Definition : Assists in the correct non-covalent assembly of polypeptide-containing structures in vivo, but is n a component of these assembled structures when they are performing their normal biological function.

Parents :
[GO:0003674] molecular_function

Children :
[GO:0003763] chaperonin ATPase activity
[GO:0003767] co-chaperone activity
[GO:0015472] fimbrium-specific chaperone activity
[GO:0003759] glycoprotein-specific chaperone activity
[GO:0003773] heat shock protein activity
[GO:0003762] histone-specific chaperone activity
[GO:0008077] Hsp70/Hsp90 organizing protein activity
[GO:0016530] metallochaperone activity
[GO:0000005] ribosomal chaperone activity
[GO:0017072] tubulin-specific chaperone activity

Synonyms :
GO.0003757
GO.0003758
GO.0003760
GO.0003761

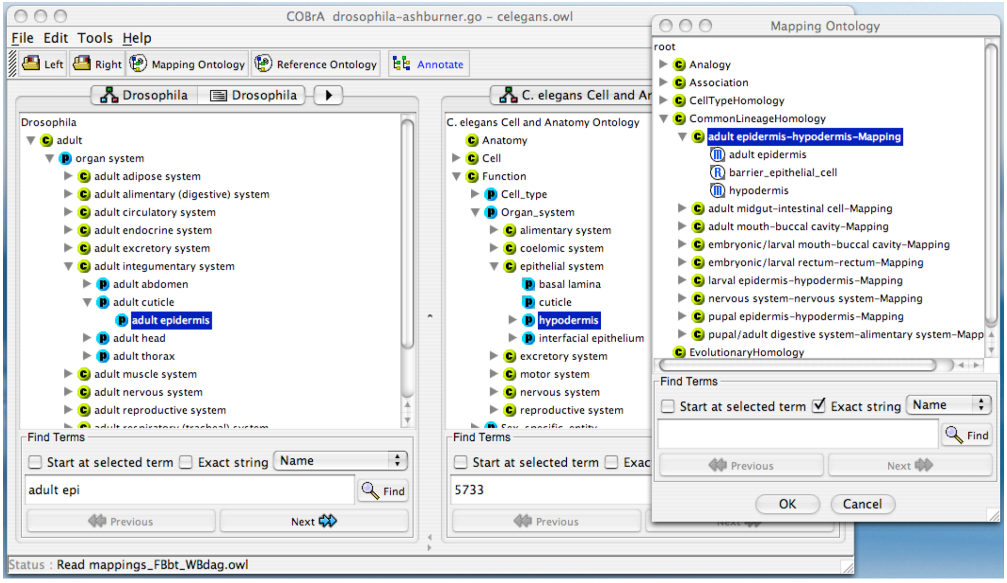
Database Cross References :

Database Symbol	DB Cross Reference	Reference Type
GenProtEC	5.7 chaperoning	
SP_KW	Chaperone	
InterPro	IPR001305	
InterPro	IPR001404	
InterPro	IPR001829	

Status : Read go-200310.xml

Fig. 1. The COBrA bio-ontology editor. The lefthand panel shows a tree view of the ontology and the righthand panel a node-view of the term highlighted on the left. Both views are navigable.

Beginning with the stand-alone functionality, COBrA-CT will be upgraded to read and write all of the current bio-ontology language syntaxes, including the Gene Ontology OBO format and the formats supported by tools such as DAG-Edit[5]. In ontology editing mode, COBrA will perform error checks to assist the user to detect cycles in the ontology and inconsistencies arising from the domain and range restrictions of relationships. The tool will also prompt the user for information on the disjointness of classes, and to create exhaustive sets of subclasses, as these are sound ontology engineering principles. Users ought to provide textual definitions for terms, but these are often omitted and COBrA will prompt users for this information. Use of this tool will be optional, as some users find such tools intrusive, but a record of whether the user has reviewed their ontology edits in this manner will be kept. This information can be used as evidence for the amount of review that an ontology has undergone and contributes to the provenance of the ontology. Curators will be aided by ensuring that proposed changes conform to minimal requirements prior to submission.



□ Start at selected term □ Exact string Name: adult epi

□ Start at selected term □ Exac: 5733

□ Start at selected term ☑ Exact string Name

Status : Read mappings_FBbt_WBdag.owl

Fig. 2. COBrA being used to create homology links between the drosophila (left) and C. elegans (right). The righthand panel shows the types of mapping.

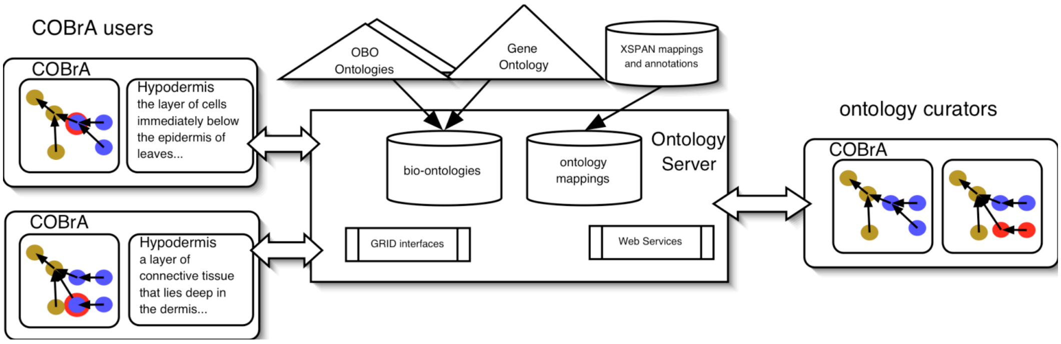


Fig. 3. The Ontology Server and its interaction with the COBrA user and ontology curator.

Ontology Management

Organising the curation effort in a distributed setting, providing access to current and past versions of ontologies and providing search and related services requires an ontology management server. While it is possible (and certainly common) to simply archive different versions of an ontology, there is much to be gained from an explicit record of the changes made and their rationale which COBrA-CT will record as mappings between ontology versions. Ontology version mappings will be considered in the curation process, as they provide the explanation for the proposed changes. The focus on bio-ontologies is important as the problems we address are very complex in the general case. However, strategies have evolved in bioinformatics to address them, for example, concepts have IDs that are unique, and rather than being deleted, IDs persist as annotations to other concepts, or are categorised as obsolete terms.

As we shall be adopting the Web Ontology Language (OWL) with its XML syntax as the means of data exchange, we shall be able to take advantage of both ontology-based and XML-based techniques for capturing changes. The difference between these can be illustrated as follows: from a structural perspective, an edit to an XML-encoded ontology that asserts that class **C**, known to be a subclass of **A,** is also a subclass of **B** would be viewed as adding an edge to the XML structure (irrespective of anything else we know about **A** and **B**). If classes **A** and **B** are known to be disjoint, then from the ontology perspective we would note a contradiction in the semantics as there can be no common subclass of disjoint classes. We propose to layer semantic checks on an XML-based ontology archiving mechanism. This approach is flexible, as XML is very widely adopted, and can exploit (but is not committed to) the logical language an ontology is expressed in.

It has been noted (Buneman, 2002) that changes to scientific data archives are accretive – most changes are additive – although deletion and modification also occur. Scientific data is typically structured hierarchically, allowing a hierarchical key structure to be exploited in archiving changes to the data. Managing versions of a data resource can be performed on the basis of *diffs* (i.e. by recording the editing steps that cause the change). However, there are advantages for an approach where all objects have an associated timestamp. The central notions of hierarchical organisation, objects and timestamps (Buneman, Fan et al. 2001, 2002) also apply to ontologies and ontology management, and this is the approach we plan to adopt. Given the problems noted by Noy et al. (2003) with the simple *diff* approach, our approach will also be structure-based. We shall identify types of ontological changes that occur in practice, taking the procedures used in practice, e.g. by the Gene Ontology, as a starting point. As we do not assume that ontologies will make use of formalisms such as Description Logic, our approach is not reliant on the widespread uptake of this particular logic. However, we will exploit any formalism that is associated with an ontology, which may be DL or first-order logic (Aitken, 2005b).

A simple model for assigning rights to users to allow them to download, upload, and publish ontologies will be defined to control the curation process. We shall also consider explicitly representing the 'process' of curation in explicit process models, e.g. from authoring, through review, to publication and revision. The ontology curator will require a visualisation of the differences between two versions of an ontology and we can provide this through COBrA's dual-view capability.

Given that the ontologies will be stored centrally, and separately from the numerous databases that store the data and its annotation, we can offer an important service by providing the managers of biological databases with the version history associated with an ontology term. A term may have become obsolete since its use in annotation. There are obsolete terms in the Gene Ontology, and in the C. Elegans anatomy ontology which has been significantly reorganised in the past year. In both cases, references to obsolete concepts persist as annotations and, should these annotations be updated or queried, the ontology server can both identify that obsolescence has occurred, and find the candidate replacement concepts by winding the ontology forward from the version used in annotation to the current version.

Related Work: COBrA is distinguished from generic ontology editors and environments such as Protégé and Prompt in that COBrA (and the COBrA-CT Protégé plug-ins) is tailored for life science uses. We note that tools that do not provide the support that users need have only a small uptake among biologists. DAGEdit, like COBrA, is designed to support the editing of bio-ontologies. However, DAGEdit does not support ontology mapping in any way, and does not fully support the Web Ontology Language. DAGEdit has only simple archiving facilities based on CVS. In contrast with DL approaches, we do not assume that all ontologies will be in OWL's Description Logic fragment, and so will be required to handle the less formally-specified bio-ontologies. Where possible, we aim to utilise these parallel efforts and not duplicate them.

References

Aitken, J.S., Webber, B.L. and Bard J.B.L. (2004) Part-of Relations in Anatomy Ontologies: A Proposal for RDFS and OWL Formalisations. *Proc. Pacific Symposium on Biocomputing* **9**:166-177.

Aitken, S., Korf, R., Webber, B. and Bard, J. (2005a) COBrA: A Bio-Ontology Editor. *Bioinformatics.* 21(6):825-826.

Aitken, S. (2005b) Formalising concepts of species, sex and developmental stage in anatomical ontologies. *Bioinformatics* in press –doi:10.1093/bioinformatics/bti409

Bard, J.B.L. (2002) Growth and death in the developing mouse kidney: signals receptors and conversations. *BioEssays*, **24**, 72-82.

Bard, J.B.L. (2005) Anatomics: the intersection of anatomy and bioinformatics. *J. Anatomy.* In press.

Bard, J.B.L and Rhee S.Y. (2004) Bio-ontologies and the linking of phenotypes to genotypes. *Nature Review Genetics*, **5**, 213-222.

Bard, J.B.L, Rhee, S.Y. and Ashburner, M. (2005) An ontology for cell types. *Genome Biol.* In press.

Buneman, P., Davidson, S., Fan, W., Hara, C. and Tan, W. (2001) Keys for XML. Proc. WWW 10:201-210.

Buneman, P., Khanna, S., Tajima, K. and Tan, W. (2002) Archiving scientific data. ACM SIGMOD:1-12.

Canevet, C., Bickmore, W. and Webber, B.L. (2004). Towards automating the curation decision for the Nuclear Protein Database. Poster A-30, *ISMB-04*.

Heymann, S. et al (2005) Enhancing the semantics of links and paths in Life Sciences Sources. *Proc. Workshop on Database Issues in Biological Databases* (DBiBD), Edinburgh.

Lord, P. and MacDonald, A. (2003) Data curation for e-Science in the UK: an audit to establish requirements for future curation and provision. JISC Report

Luger, S. and Aitken, J.S. (2004) Cross-species Mapping between Anatomical Ontologies Based on Lexico-syntactic Properties. Poster C-40, *ISMB-04*.

Noy, N., and Musen, M. (2003) The PROMPT suit: Interactive tools for ontology merging and mapping. *International Journal of Human-Computer Studies*, 59/6:983-1024.

Parkinson H. et al. (2004) The SOFG Anatomy Entry List (SAEL): an annotation tool for functional genomics data. *Bioinformatics.* In press

Smith B, Williams J, Schulze-Kremer S. (2003) The ontology of the gene ontology. *Proc. AMIA Annual Symp 2003*:609-13