BIOINFORMATICS APPLICATIONS NOTE

Vol. 28 no. 6 2012, pages 895–896
doi:10.1093/bioinformatics/bts029

Databases and ontologies

Advance Access publication January 16, 2012

CODEX: exploration of semantic changes between ontology versions

Michael Hartung[1,2,*], Anika Gross[1,2] and Erhard Rahm[1,2]

[1]Interdisciplinary Center for Bioinformatics, University of Leipzig, Leipzig, Germany and [2]Department of Computer Science, University of Leipzig, Leipzig, Germany

Associate Editor: Alex Bateman

ABSTRACT

Summary: Life science ontologies substantially change over time to meet the requirements of their users and to include the newest domain knowledge. Thus, an important task is to know what has been modified between two versions of an ontology (*diff*). This diff should contain all performed changes as compact and understandable as possible. We present CODEX (Complex Ontology Diff Explorer), a tool that allows determining semantic changes between two versions of an ontology, which users can interactively analyze in multiple ways.

Availability and implementation: CODEX is available under `http://www.izbi.de/codex` and is supported by all major browsers. It is implemented in Java based on Google Web Toolkit technology. Additionally, users can access a web service interface to use the diff functionality in their applications and analyses.

Contact: hartung@izbi.uni-leipzig.de

Received on November 25, 2011; revised on January 4, 2012; accepted on January 11, 2012

1 INTRODUCTION

Recently ontologies have become very popular in the life sciences. They consist of a harmonized vocabulary of terms (concepts) describing and structuring a domain of interest. Their main application is the consistent and uniform description (annotation) of biological entities such as genes and proteins enabling analyses such as term enrichment or gene expression data studies. The most used ontology in bioinformatics is the Gene Ontology (GO) (Harris *et al.*, 2004) consisting of subontologies for molecular functions, biological processes and cellular components. A huge number of life science ontologies are managed within the Open Biomedical Ontologies Foundry (OBO) (Smith *et al.*, 2007), which has established the common OBO format as representation language.

The content of ontologies is not static. Due to new knowledge (e.g. from experimental results) or design errors made in earlier versions, they underlie continuous changes to enhance their quality. The OnEX tool (Hartung *et al.*, 2009) already allows studying the evolution of life science ontologies over the recent years, e.g. one may notice that the Biological Processes of GO doubled in their size to ~21 000 concepts between 2006 and 2011. However, OnEX considers rather simple changes (e.g. concept additions and deletions) such that there is still an increasing need to figure out what has exactly been changed between the two versions of an ontology (diff). The complexity and size of life science ontologies requires that a diff should be as compact and human-understandable as possible. Semantic changes such as concept merges, additions of entire subgraphs or moves of concepts should be reported instead of long lists of rather basic changes as generated by tools such as *obodiff* of OBO-Edit (Day-Richter *et al.*, 2007). A semantic diff can be valuable for both ontology users and developers, e.g. if a user likes to perform a term enrichment analysis based on a new ontology version, she may be interested in revised concepts compared with the previously used version, or a developer likes to know which changes were performed in the last year to plan future modifications. Current ontology browsers such as AmiGO or Ontology Lookup Service are limited to the latest available version and do not offer diff facilities. We therefore present CODEX, a web tool to *ad hoc* compute a semantically rich diff between two ontology versions.

*To whom correspondence should be addressed.

2 OVERVIEW OF CODEX

CODEX consists of a web front-end, a web service interface and a repository for diff calculations at the back-end. The main diff algorithm used by CODEX is described in Hartung *et al.* (2010). Particularly, it first performs a match between the input versions to determine equal or slightly changed ontology elements. It then generates a basic diff which is based on simple changes: *add*, *del* and *map* of concepts, relationships and attributes. Taking this basic diff as input, an iterative application of rules generates more and more semantic changes (e.g. *merge*, *split*, *substitute*, *move*, *addLeaf*, *addSubGraph*, *toObsolete*, …) as long as a compaction is possible. Thus, the final diff result represents the most compact diff between the input versions. For a full list and description of supported changes, we refer to Hartung *et al.* (2010) or the help pages of CODEX. We will now describe the web front-end and web service of CODEX in more detail and then show some application results from determining diffs for selected ontology versions.

2.1 Web front-end

The web front-end contains two main sections as displayed in Figure 1. The *input section* offers different possibilities to specify the input versions for a diff computation. CODEX supports ontologies formatted in OBO as well as OWL. First, users can directly put the ontology contents into appropriate text areas (A1). Second, one can simply provide the URLs of the ontology versions to process (A2). Third, there is the possibility to upload ontology files stored on the local machine of a user (A3). Finally, ontology versions already available in OnEX can be selected as input (A4). Currently, OnEX provides access to ~850 versions of 16 ontologies. For each variant, a submit button is used to start the diff computation after the input has been specified. Example buttons allow for running predefined examples.

After the online diff computation, users can enter the *result analysis section* to explore the resulting diff in more detail. For a first overview, one can study the *Overall Statistics* (B1) of the ontology versions and their diff. Version statistics such as number of concepts and relationships, the

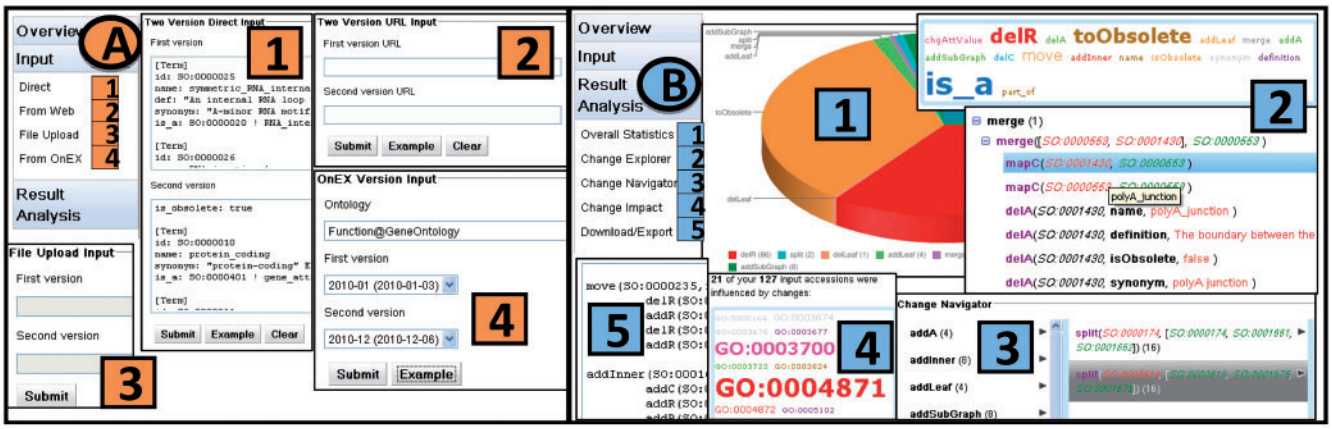

Fig. 1. Web front-end of CODEX. The *input section* **(A)** offers different input possibilities: (A1) copy/paste ontology contents, (A2) specify web URLs, (A3) upload of ontology files or (A4) use versions of OnEX. In the *result analysis section* **(B)**, users can analyze the diff result in multiple ways: (B1) Overall Statistics, (B2) Change Explorer, (B3) Change Navigator, (B4) Change Impact and (B5) Download/Export.

diff sizes as well as growth rates are presented in a table. Furthermore, the content of the diff can be analyzed via pie charts (with different selection options) such that users can recognize the occurred changes and their frequencies.

The *Change Explorer* (B2) utilizes tag clouds to present the most frequent changes or elements that have been influenced by a change. If users click on the displayed terms, they can explore the corresponding changes in a tree-like manner. The root nodes show the change type and its frequency. For each change one can drill down into its details, i.e. one can inspect the basic changes that belong to a complex change, e.g. all concept and relationship additions involved in a subgraph addition. A similar functionality is offered by the *Change Navigator* (B3) where one can flexibly drill through the compact diff starting at the most compact changes.

The *Change Impact Analysis* (B4) offers the possibility to find out which concepts of a given set of accession numbers of interest were influenced by a change or not. Again tag clouds and tree-based navigation simplify the exploration of changes. Finally, *Download/Export* (B5) can be used to copy/paste results of the diff for further processing. On the one hand, it is possible to access the overall diff statistics, e.g. for generating summaries or charts. On the other hand, all determined changes are presented in text format and can thus be reused in further application and analysis scenarios.

2.2 Web Service

In addition to the web front-end application, programmers are able to access the diff functionality via a web service interface available at http://dbserv2.informatik.uni-leipzig.de:4478/contodiff_ws?WSDL. Using the provided WSDL description, they can generate corresponding client classes that enable web service interaction. Particularly, the service accepts OBO, OWL or OnEX ontology versions as input and generates the compact as well as the basic diff as output.

3 APPLICATION

We apply CODEX to determine and analyze the difference between selected ontology versions. First, we compute the diff between the July and November 2010 version of the Sequence Ontology (Eilbeck *et al.*, 2005) [see example in (A1)]. There is an increase of about 2% and the resulting diff contains 260 changes whereby semantic changes such as *toObsolete* (81) and *move* (54) dominate [see tag cloud in (B2)]. This shows that a lot of knowledge has been revised. In contrast, eight subgraphs have been inserted; the largest one (SO:0001659—promoter_element) contained 11 concepts.

As a second example, we determine the diff between the January and December 2010 version of GO Molecular Functions [see example in (A4)]. This time the ontology grew by ∼3% and the diff encompasses 3300 changes. The changes are dominated by attribute value changes (2423) especially changes of definitions. Other frequent semantic changes are moves of concepts (188) and leaf additions (169). However, there was a huge amount of revisions including merges (60) as well as obsolete changes (20). The largest of the 33 added subgraphs with 29 new concepts is GO:0000988 (protein binding transcription factor activity).

4 CONCLUSION

The presented CODEX application allows to compare different ontology versions and determines a compact diff based on semantic changes. Thus, users can flexibly determine semantic diffs and use this knowledge for understanding the ontology evolution or to adapt dependent data. The application can be accessed via an interactive web front-end or a web service interface.

Funding: DFG grant [RA 497/18-1] (Evolution of Ontologies and Mappings).

Conflict of Interest: none declared.

REFERENCES

Day-Richter,J. *et al.* (2007) OBO-Edit–an ontology editor for biologists. *Bioinformatics*, **23**, 2198–2200.

Eilbeck,K. *et al.* (2005) The sequence ontology: a tool for the unification of genome annotations. *Genome Biol.*, **6**, R44.

Harris,M.A. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, 258–261.

Hartung,M. *et al.* (2009) OnEX: Exploring changes in life science ontologies. *BMC Bioinformatics*, **10**, 250.

Hartung,M. *et al.* (2010) Rule-based Generation of Diff Evolution Mappings between Ontology Versions. *CoRR*, **abs/1010.0122**.

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