Nature Precedings : doi:10.1038/npre.2010.4585.1 : Posted 28 Jun 2010

CDAO-Store: A New Vision for Data Integration

Brandon Chisham*, Trung Le, Enrico Pontelli, Tran Son, Ben Wright†
Department of Computer Science
New Mexico State University
{bchisham,tle,epontell,tson,bwright}@cs.nmsu.edu

The *Comparative Data Analysis Ontology (CDAO)*[1] is an ontology developed, as part of the EvoInfo[2] and EvoIO[3] groups supported by NESCent[4], to provide semantics to the descriptions of data and transformations commonly found in the domain of phylogenetic inference. The core concepts of the ontology enables the description of phylogenetic trees and associated character data matrices.

CDAO-store is a repository providing a rich set of API's for querying and visualizing phyloinformatics data. It is a triple-store encoding data as RDF triples constructed according to the CDAO concept vocabulary. CDAO-store provides three classes of services: a service for importing data in CDAO format, a PhyloWS interface supporting an advanced set of queries for other external applications, and a web-interface for visualizing and interacting with data in the store.

The import feature is quite flexible and allows importing not only raw data, but also arbitrary annotations. Currently, we have imported all of the trees in the TreeBASE dump dated January 2009. The corresponding annotations relating these trees to studies and their authors have also been imported. We also have a translation service to convert Phylip, NEXUS, MEGA, and NeXML format data into CDAO format for the import service.

The store provides a PhyloWS interface for programmatic access to data. The interface supports retrieving trees by id, finding the nearest common ancestor of a set of taxa in a tree, or finding the minimum spanning clade for a set of taxa. Results are returned as RDF/XML format CDAO documents. Additionally, the interface supports retrieving matrices. The interface also supports extracting lists of trees that match certain structural criteria such as the number of nodes, leaves, internal nodes, the radius, or diameter of the tree.

We have also implemented visualization tools, called CDAO-Explorer, for both trees and matrices. This collection of tools allow users to find, view, and interact with data in the store. Once a dataset is selected one can view its structure, and interact with it, by adding annotations or customizing their view of it. For example, users can select particular columns and rows from a matrix to view more closely only part of it. Users can also view a matrix as a color-coded figure. With trees, users can view two different layouts. One is a node layout and the second is a layout based on dynamic positioning. A user can also search the tree visualization for node or edge name. Also, additional information on individual nodes or edges is available. This last feature is currently under development. The tree visualization is based on the prefuse framework found at `http://prefuse.org/download/`

The web-interface supports searching for trees by TreeBASE accession number, phylogenetic method, construction algorithm, study, author, or taxonomic identifier. In addition, it supports display of basic query results on the web such as finding the nearest common ancestor or the minimum spanning clade of a set of nodes in a tree. It also supports viewing lists of trees of a particular size. We have also integrated the CDAO-Explorer set of tools into the web-site, so that one may visualize a particular tree that one has found as the result of a query.

The CDAO-store is available at `http://www.cs.nmsu.edu/~cdaostore/`, the tool set including the translator is licensed under the GPL, and is available on sourceforge `http://cdaotools.sourceforge.net/`.

*Presenter
†Presenter
[1] `www.evolutionaryontology.org`
[2] `https://www.nescent.org/wg_evoinfo/Main_Page`
[3] `http://evoio.org/wiki/Main_Page`
[4] `http://www.nescent.org/index.php`