CUAHSI

universities allied for water research

HYDROTAGGER

Functional Description Document
(Version 1.0)

A guide through of the underlying technologies for the semantic tagging application HydroTagger

May 2008

Prepared by:
Michael Piasecki

Department of Civil, Architectural & Environmental Engineering
Drexel University
Philadelphia, PA 19104

Distribution

Table of Contents

1 INTRODUCTION

The HydroTagger application was developed as a support application to aid the semantic tagging of variable codes originating in WaterOneFlow Web Services and the search ontology concepts. The tagging is a necessary step for the search engine HydroSeek (reference?) to be able to find the underlying datasets when queried for a keyword and temporal bracket. The key objective of this application is to supply a graphical user interface (GUI) that is easy to use and intuitive enough so non-experts can easily find and tag (or map) the variables of the Web Service they administer to the leaf level concepts in the search ontology.

Because the search ontology, despite just covering a relatively small horizon of possible variables and parameters, is already fairly large (as of April 2008 there are about 250 concepts in the ontology), visualization is a critical aspect that needs to be addressed. This is because current ontology visualization tools are largely inadequate to traverse ontologies larger than 50 entries in a fashion that is intuitive and easy. Visual aids in traversing the ontology is crucial when trying to grasp the concept branches so the individual carrying out the mapping can "see" where the proper path and location for the variable is at the leaf level. This issue has posed a special problem and challenge that was met at the San Diego Supercomputer Center by purchasing a license for a so-called hyperbolic Star Tree viewer. This visualization approach is the only currently acceptable way of traversing large concept constructs given the limited viewing plane that a normal screen offers to a user. The application is installed at the San Diego Supercomputer Center and is embedded into the HydroTagger application as an applet that is loaded when invoking the HydroTagger.

The HydroTagger is intended to help perform the mappings but also offers the opportunity to suggest new concepts at the leaf level in case appropriate ones cannot be found. This feature, albeit limited, was intended to give the user community a first tool to actually expand and build out the search ontology. This aspect was deemed important to the underlying functionality of the HydroSeek environment so as not to force users and data managers to use concepts that may not work for them.

The HydroTagger application works together with a Catalog Harvester program that trawls through a list of specified WaterOneFlow Web Services at regular intervals. This application, which is a standalone Windows application, can be scheduled at desired intervals to "sniff" out the latest additions to a Web Service's underlying database via the Windows Task Scheduler. The Catalog Harvester makes use of the regular WaterOneFlow Web Service methods to identify new variable codes that have been added during the last update period and offer these up for mapping in the HydroTagger interface. To this end the Observations Data Model (ODM) design has been amended by a number of tables that keep track of unmapped variables, new concepts, and in particular the concept code ⇔ variable code mappings that provide the link between variables and search keywords.

This document seeks to outline the functionality requirements that support the HydroTagger tagging environment as well as to define key technologies used in this approach, namely the knowledge base comprised of a collection of layered ontologies that support the keyword search and some specialized tables added to the ODM.

2 HydroTagger Design

The HydroTagger is a support application to the HydroSeek environment. Its sole purpose is to present end users (or data managers) a simple to use tool to manually execute the mapping between the variable codes available through their respective Web Services and the search ontology, i.e., the tagging to a search concept. In order to facilitate "ease-of-use" the application features a GUI (see Figure 1) that displays the trees and leafs of the search ontology (top panel) and then displays all newly detected variable codes in a small table together with the mapping section and a section showing the executed mappings in the bottom panel (blue shaded).

2.1 Graphical User Interface

The graphical display is based on a hyperbolic Star Tree viewer from Inxight, a license for which has been purchased and hosted by SDSC. The Star Tree viewer is the only visualization approach currently available that allows a reasonable viewing and traversing of taxonomies (or ontologies) that have more than about 100 members. Hence, the Inxight license is vital to the functioning and supporting role of the HydroTagger. It should be noted that the Star Tree viewer can only display purely hierarchical tree structures, i.e., taxonomies, hence it is not ideally suited for ontology display because of the multiple parent situation a class might encounter. This is solved by simply duplicating the class so it can appear in all places as demanded by the ontology. Also, the Star Tree viewer does not accept ontologies (or OWL files, precisely for the reason just mentioned), hence it requires an auxiliary code that converts OWL files into Inxight Star Tree Viewer format.

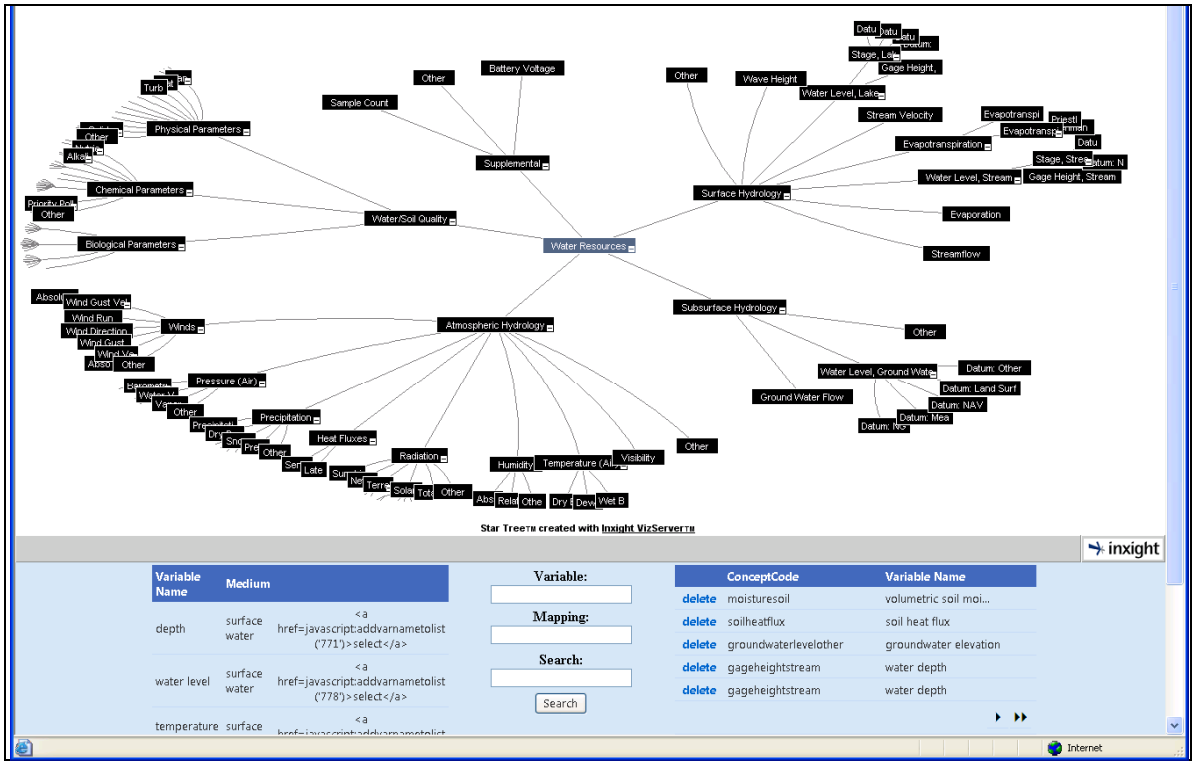


Figure 1 HydroTagger Graphical User Interface

The left side of the lower panel actually displays variable name and the sample medium the variable was collected in. It should be noted that the display options do not reflect the unique ID the mapping is executed on. Rather, Variable Name and Medium are just for identification purposes while the mapping is executed on a unique variable ID assigned to this variable code by the HydroSeek ODM whenever a new variable code has been detected in any of the registered Web Services. Hence, the displayed mappings (right side of lower panel) show a concept code ⇔ variable name paring that really consists of the concept code ⇔ variable ID mapping that is associated with the variable name.

The user interface also permits a reverse action, i.e., the suggestion of new concepts by data managers in case they cannot find an adequate concept to map their variable to. This in effect allows the user to build and extend the ontology and installs the possibility of user community involvement in developing the search ontology without having to know OWL, or ontology editors like Protégé or SWOOP. When suggesting a new concept the suggestion is stored in the MappingsPending table for review by the ontology keepers. If the suggestion is approved the ontology is automatically updated, and the Star Tree viewer is able to display the added concept shortly after (the OWL to Inxight conversion code still needs to run once so the viewer base file is updated before being displayed). Currently, new concepts can only be suggested at leaf levels, i.e., a concept that can be mapped to. This "horizontal" only addition is currently a severe limitation of the HydroTagger that effectively renders it an inadequate tool for community ontology development.

2.2 TABLES SUPPORTING HYDROTAGGER

The HydroSeek ODM was modified to accommodate an additional set of tables in support of the HydroSeek and HydroTagger functionalities, as discussed in "HydroSeek_Functional_Description" (reference). Four of the table additions are in direct support of the HydroTagger application and act as a repository for the mappings that have been carried out or need to be carried out as a result of the updating service that is invoked frequently. These four additions (some of which are shown in Figure 2 below) are those of the

- NONTaggedVariables table that holds all those variable codes that have been identified as not-yet-present in the HydroSeek system. These entries are the result of a regular (currently once a week) trawling action by a program called the Catalog Harvester that interrogates all (or a subset of) registered WaterOneFlow Web Services.
- ApprovedMappings table that holds all variable codes that have been mapped to a specific concept code. Each variable code assigned is unique to the HydroSeek database and is in fact the only way HydroSeek tracks new and existing variable codes in the database. The table also contains who approved the mapping, when it was done, and what ontology version it belongs to.
- MappingsPending table that stores the new concept suggestions made by data managers. The table will accept a new concept code, a description of what it is, and also a concept name (which is displayed in the viewer). It also tracks the person who suggested it and then the date when it was approved.
- FrequentUpdates table that contains information about all those Web Service URLs that the update program is supposed to interrogate. This is not necessarily the total number of all Web Services, as some of them are too large for a weekly trawling action (like USGS NWIS and EPA STORET) or do not undergo frequent updates. The table also contains information about the URL to contact, and when information concerning this specific Web Service was last updated.

These tables are updated and accessed each time the updating service trawls through the Web Services or data managers access the HydroTagger to perform the mappings of the new variable codes that their Web Services have yielded during the last scheduled updating.

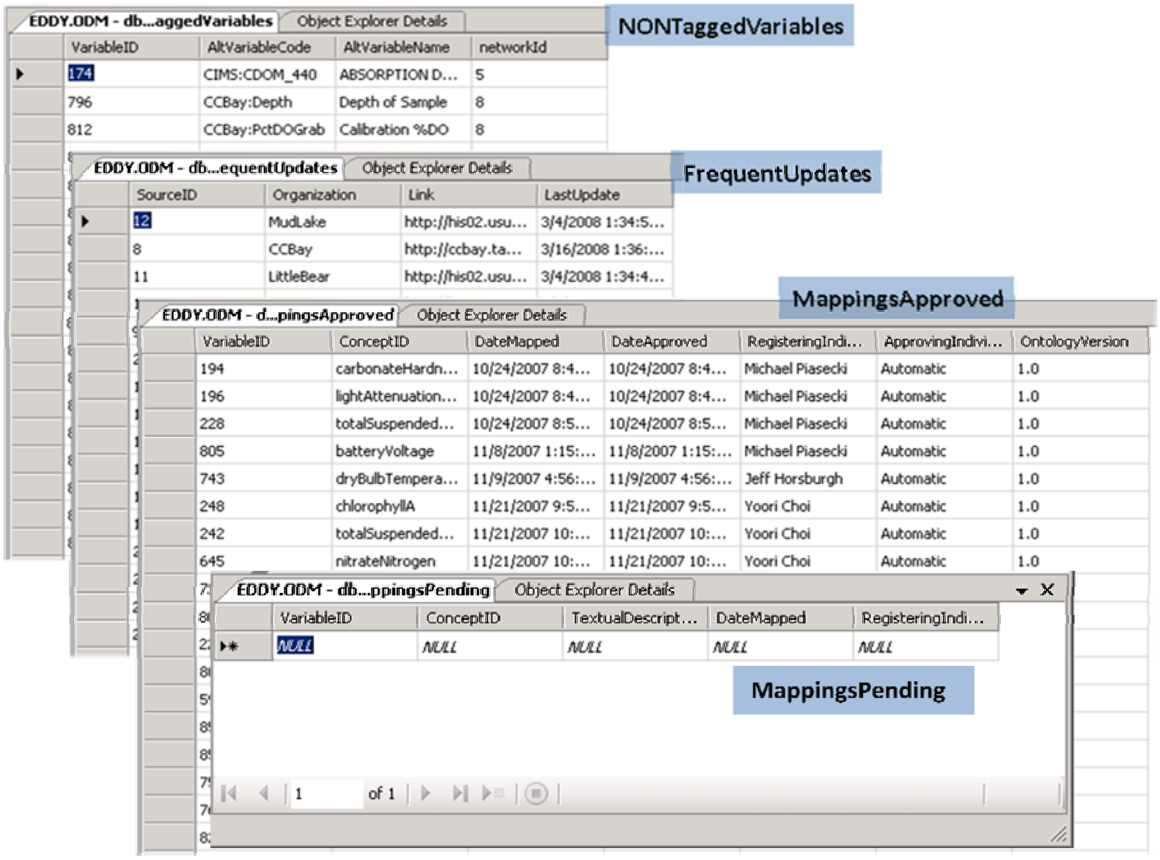
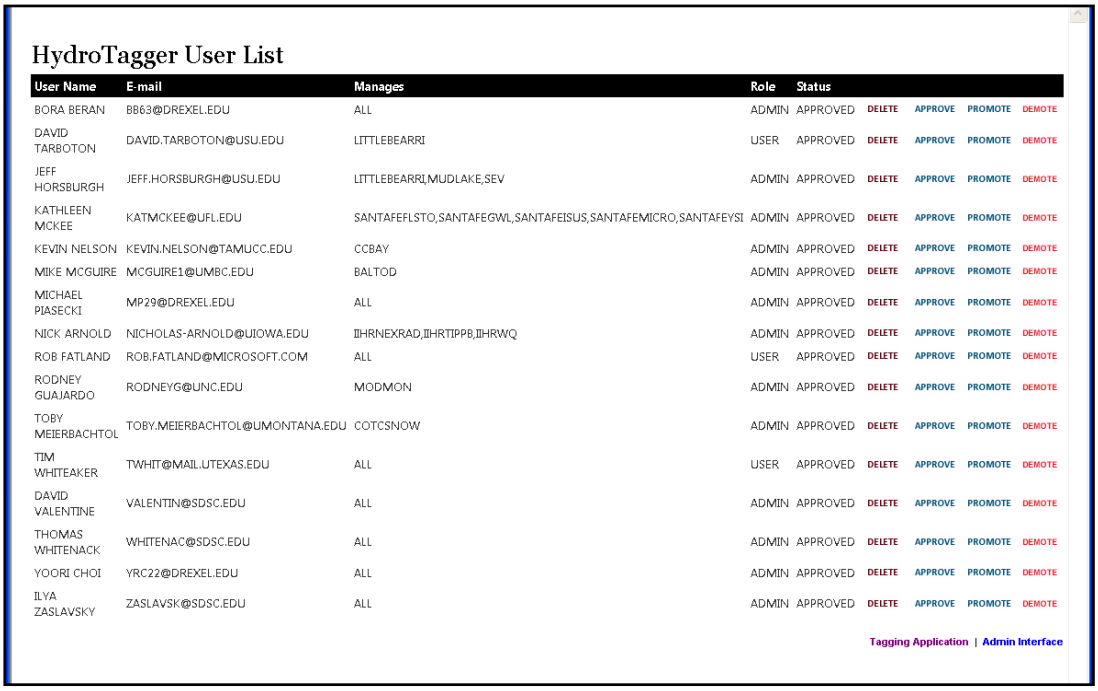


Figure 2 New Tables added to ODM to support HydroSeek

2.3 HydroTagger Administration

The HydroTagger is currently restricted to approved users. This is to minimize the administrative burden for each Web Service (only one person should carry out the mappings) and also to minimize impact on ontology updates. The system currently supports three types of users: an Administrator that manages and has permissions for "ALL", a Web Service specific Administrator who only has access to the Web Services registered in her/his name, and a Web Service specific user (to be approved by the Web Service Administrator), who can also execute mappings and suggest new concepts (see Figure 3 of the current user list, April 2008). This user list is supported by a small database that holds all relevant information and is also accessed (in addition to the HydroSeek ODM) during invocation of the HydroTagger.



HydroTagger User List

User Name	E-mail	Manages	Role	Status				
BORA BERAN	BB63@DREXEL.EDU	ALL	ADMIN	APPROVED	DELETE	APPROVE	PROMOTE	DEMOTE
DAVID TARBOTON	DAVID.TARBOTON@USU.EDU	LITTLEBEARRI	USER	APPROVED	DELETE	APPROVE	PROMOTE	DEMOTE
JEFF HORSBURGH	JEFF.HORSBURGH@USU.EDU	LITTLEBEARRI,MUDLAKE,SEV	ADMIN	APPROVED	DELETE	APPROVE	PROMOTE	DEMOTE
KATHLEEN MCKEE	KATMCKEE@UFL.EDU	SANTAFEFLSTO,SANTAFEGWL,SANTAFEISUS,SANTAFEMICRO,SANTAFEYSI	ADMIN	APPROVED	DELETE	APPROVE	PROMOTE	DEMOTE
KEVIN NELSON	KEVIN.NELSON@TAMUCC.EDU	CCBAY	ADMIN	APPROVED	DELETE	APPROVE	PROMOTE	DEMOTE
MIKE MCGUIRE	MCGUIRE1@UMBC.EDU	BALTOD	ADMIN	APPROVED	DELETE	APPROVE	PROMOTE	DEMOTE
MICHAEL PIASECKI	MP29@DREXEL.EDU	ALL	ADMIN	APPROVED	DELETE	APPROVE	PROMOTE	DEMOTE
NICK ARNOLD	NICHOLAS-ARNOLD@UIOWA.EDU	IIHRNEXRAD,IIHRTIPPB,IIHRWQ	ADMIN	APPROVED	DELETE	APPROVE	PROMOTE	DEMOTE
ROB FATLAND	ROB.FATLAND@MICROSOFT.COM	ALL	USER	APPROVED	DELETE	APPROVE	PROMOTE	DEMOTE
RODNEY GUAJARDO	RODNEYG@UNC.EDU	MODMON	ADMIN	APPROVED	DELETE	APPROVE	PROMOTE	DEMOTE
TOBY MEIERBACHTOL	TOBY.MEIERBACHTOL@UMONTANA.EDU	COTCSNOW	ADMIN	APPROVED	DELETE	APPROVE	PROMOTE	DEMOTE
TIM WHITEAKER	TWHIT@MAIL.UTEXAS.EDU	ALL	USER	APPROVED	DELETE	APPROVE	PROMOTE	DEMOTE
DAVID VALENTINE	VALENTIN@SDSC.EDU	ALL	ADMIN	APPROVED	DELETE	APPROVE	PROMOTE	DEMOTE
THOMAS WHITENACK	WHITENAC@SDSC.EDU	ALL	ADMIN	APPROVED	DELETE	APPROVE	PROMOTE	DEMOTE
YOORI CHOI	YRC22@DREXEL.EDU	ALL	ADMIN	APPROVED	DELETE	APPROVE	PROMOTE	DEMOTE
ILYA ZASLAVSKY	ZASLAVSK@SDSC.EDU	ALL	ADMIN	APPROVED	DELETE	APPROVE	PROMOTE	DEMOTE

Tagging Application | Admin Interface

Figure 3 HydroTagger User List as of April 2008

The homepage for HydroTagger is a simple HTML page (see Figure 4), that in turn offers a number of Active Server Page Extended (aspx) links that are executed on the web-server (in this case at SDSC).

Figure 4 HydroTagger Homepage

The system is divided into three sections: the first is for public “anyone” use and permits the user to map variables and new concepts at will. There are no restrictions on this interface because the underlying database is a dummy database that is not connected to the HydroSeek system. This section is intended for demo and training only.

The second section is the restricted section of the HydroTagger and only approved users can enter the system. Contrary to the first section, this section is connected to the HydroSeek system thus enabling the registered users to execute mappings and suggest new concepts. There is also a User Management Console which is for the System Administrators only. Here new user registrations are approved and individuals can be upgraded or downgraded in their user role.

The last section contains the registration site, where new users can register after having submitted their name, email address, password, and the Web Services they wish to be registered for. The request for registration is entered into the HydroTagger database and displayed in the user list whenever a System Administrator logs on.

Once the new user has been approved the database is updated, i.e., certain Web Service IDs are linked to specific users. The last item contains a short Help section that guides the user through the specific steps when attempting to map variables and suggests new concepts.

3 AUTOMATIC UPDATING: "CATALOG HARVESTER"

In order to ensure that HydroSeek is continually updated with the latest data additions that happened in any of the registered Web Services, an automated updating application called the Catalog Harvester has been developed. This application makes use of the standard WaterOneFlow Web Services that are published for each observations network and essentially trawls through the data holdings of each network by repeatedly invoking the Web Services. The Catalog Harvester is a standalone application and can be scheduled at any desired frequency in the Task Scheduler program of the operating system. Currently, the Catalog Harvester is activated once a week (Saturday night).

Upon activation the Catalog Harvester accesses the FrequentUpdates table in the HydroSeek ODM to find out what Web Services are to be interrogated for the current trawl. Note that the FrequentUpdates table is not identical with the Source table because certain Web Services may have been taken out of the updating sequence. For example, the three USGS Web Services are not typically trawled because the catalogs (hosted at SDSC) are not updated on a regular basis. The same applies to other nationwide data sources like EPA STORET and SNOTEL. This currently limits the range of the Catalog Harvester to local installations of the ODM distributed at the test bed sites or those sites that have added themselves to the HIS Central registry (like the IdahoWaters node). This also allows control over what Web Services in the test beds should not be interrogated because the test bed manager of that Web Service may need to conduct some maintenance on the datasets, for example QA/QC before the data can be released.

The Catalog Harvester works through the Web Service list to investigate what new holdings are available. It has two objectives:

- Find all variable codes used in that Web Service and compare it to the list of variable codes that are known to the HydroSeek tables. Those that are new are placed in the NONTaggedVariables table waiting to be mapped while the data associated with the new variables are stored in the ODM. Note that because the mapping has not been carried out, the new data will not be accessible in the HydroSeek system.

- Identify all new data additions that have been added during the last updating period. For those variable codes that are known the Catalog Harvester compares the time coverage extracted from the latest trawl with that of the HydroSeek catalog. If there is new data, the Catalog Harvester places the new data into the HydroSeek ODM thus updating the catalog holdings.

The Catalog Harvester while hosted on the WATER server at SDSC, currently only updates the HydroSeek.org installation (hosted at Drexel's EDDY server) but not HydroSeek.net (hosted at SDSC) or any other derivative or copy made of the HydroSeek.net installation. Ideally, the two underlying HydroSeek database instances, i.e., the one on WATER and the one on EDDY, should be synchronized in order to provide a complete view of the HydroSeek catalog holdings.

4 FUTURE OUTLOOK

The HydroTagger in its current stage serves well the purpose it was designed for, i.e., a tool that helps the data managers to map variable codes to search concepts. However, while the basic functionality has been established, its functioning within the WATERS network is not fully worked out both in terms of the role it plays in updating catalogs as well as a tool to expand and develop the search ontologies or, going even further, a general tool to build any type of ontology for the WATERS community. Hence, there are three main thrust areas of further development and improvement:

- Firstly, the tool (and this includes the Catalog Harvester) needs to be consolidated in one place and in one seamless application pool. Because of the manual labor that needs to go into the tagging, only a single HydroSeek ODM ought to be the repository for all mapping actions and Catalog Harvester updates. This central master database can then be synchronized with other fallback databases. This would ensure that all HydroSeek databases are in sync at all times. The tool also needs to be integrated with the HIS Central registry so there is an increased degree of automation.

- Secondly, the Tagger interface needs to be upgraded and redesigned to include more information. First responses from the user community (test bed data managers) have indicated that not enough information is on display in the tagging pane to always fully identify (by name and medium) what variable code is to be mapped. The panel may need to display more information or provide a feature that displays as a pull out menu a selected set of metadata associated with the variable code to quickly and unambiguously identify the variable code in question. The functionality should be increased to permit more flexibility in adding concepts both in the horizontal and vertical directions. This would ensure more opportunity for variable codes to be placed appropriately in case concepts are not available.

- Thirdly, the Tagger can be developed to be an ontology development tool for the community. The objective is to avoid the usual ontology tools that require a steep learning curve and in general lack adequate visualization capabilities to actually “see” the concept framework in its entirety. Visualization is a key concept when developing ontologies because of the many entries an ontology can acquire eventually becoming an unwieldy conglomeration of text strings that are very difficult to follow and comprehend. Hence, the objective is to use the visualization interface (the best on the market) as a means to access and edit the underlying ontology by adding simple editing functions like deleting, clipping and re-attaching class connections, text editing, new concepts, merging and so on. If the demands on the use of OWL remain relatively simple (for example one may just want to allow OWL-lite as a start) then a tool of this capability would be a novel contribution to the field of ontology and at the same time permit the WATERS community to actively build ontologies without having to learn a complex ontology language together with the editing tools currently available.