The Terminizer: An interactive web-based tool for the automated detection of ontological terms in unstructured, free-text annotation.

David Hancock[1,3], Norman Morrison[1,3], Giles Velarde[2], Dawn Field[3]

1. Dept. Of Computer Science, The University Of Manchester, Oxford Rd., Manchester, UK 2. Manchester Interdisciplinary Biocentre, 131 Princess Street, Manchester, UK 3. NERC Environmental Bioinformatics Centre, CEH Oxford, Oxford, UK

NERC Environmental Bioinformatics Centre

NATURAL ENVIRONMENT RESEARCH COUNCIL

The proliferation in the amount of data available online has given rise to many new challenges for those wishing to exploit it in their research. Simply finding the data can be a problem. The sheer number and diversity of sources makes it very difficult for an individual to be aware of all of the possible data sets that might be of interest to them. A second significant issue is that of data-integration. Successful integration requires that unambiguous meta-data descriptions are available to ensure that disparate data sets are comparable.

An important development which promises to help in both of these cases is the semantic web and the attendant rise in interest in ontologies. Ontologies offer the potential to assist in both the *searching* for (by enabling smarter matching and automatic generation of search terms) and the *interpretation* of data sets (where the unambiguous nature of ontological annotation facilitates the discovery of suitable mappings from one data set to another).

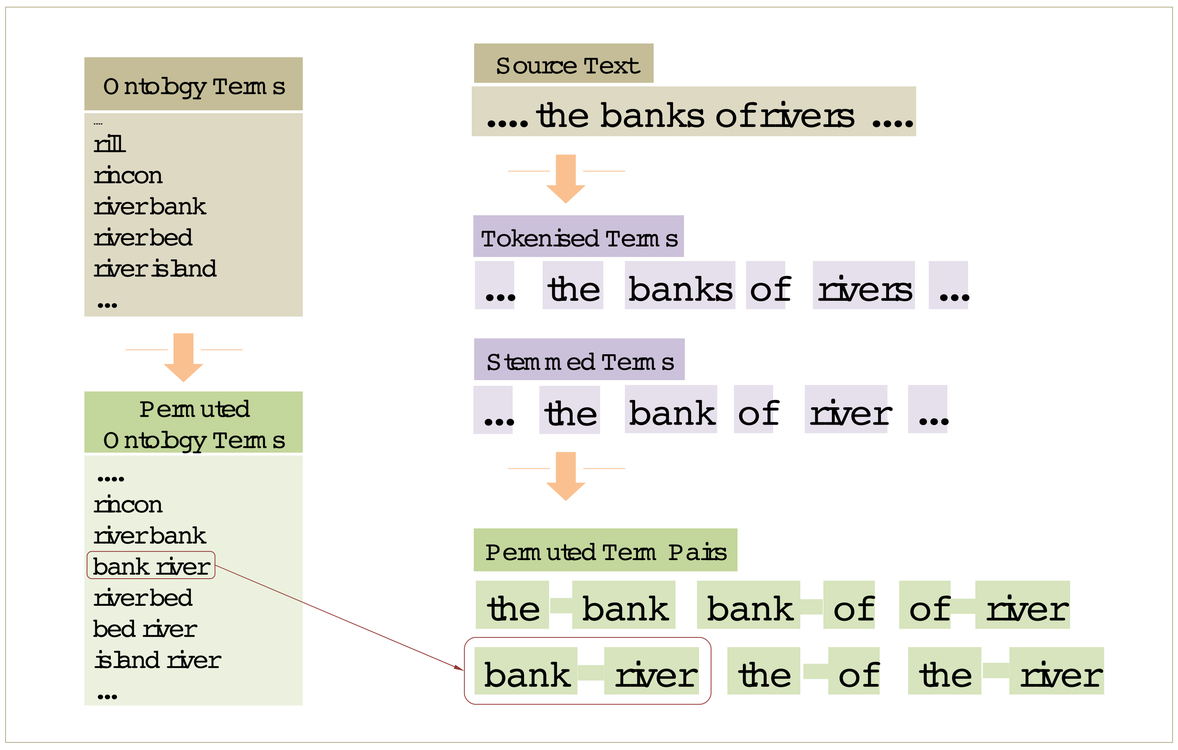

Figure 1 – Term matching

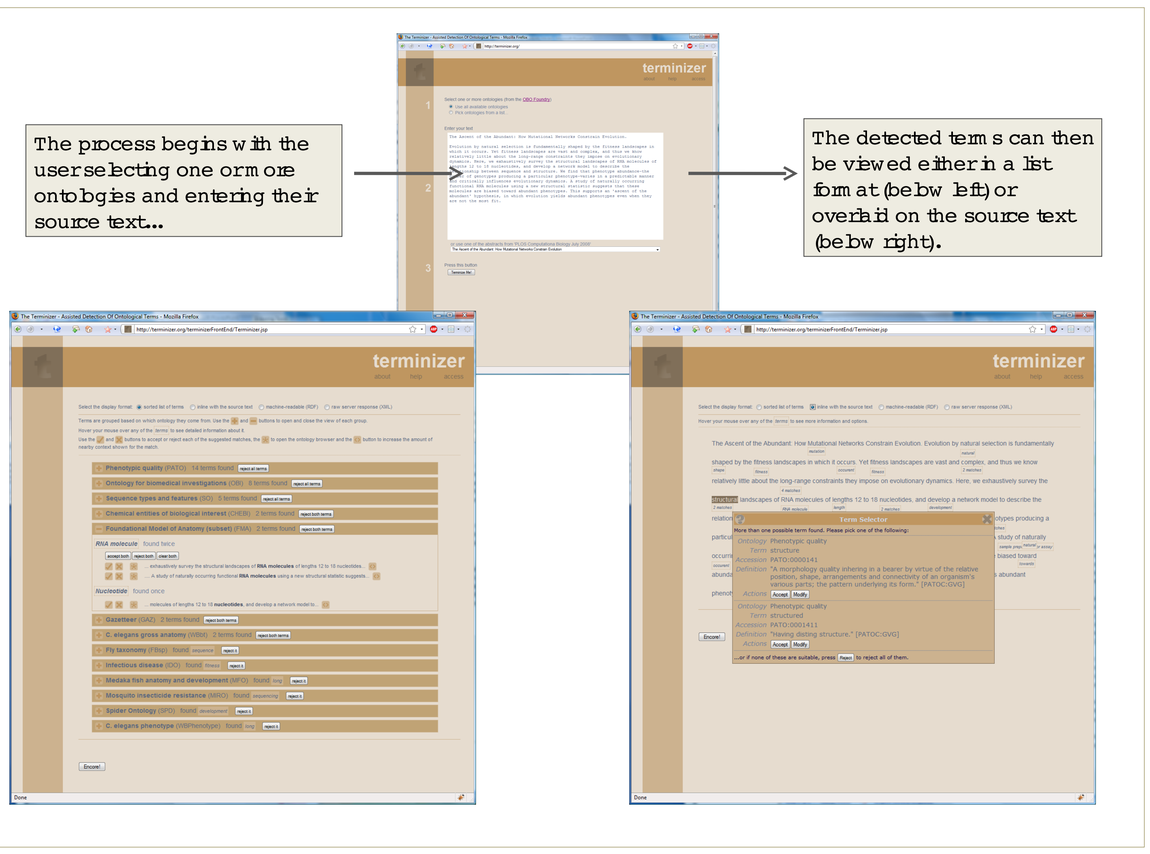

Figure 2 – The browser-based interface

The Terminizer services currently is built using the ontologies from the OBO Foundry, a collection of over 40 biological ontologies in a common format. Coupled with the GAZ gazetteer, the database presently contains 390,000 terms and 150,000 synonyms. We will shortly be expanding this to include ontologies from the National Center for Biomedical Ontology.

In addition to the interactive mode, the software is also available as a Web service. Both the term detection service and the interactive presentation layer can be incorporated within other Web sites or programs.

The Terminizer system has been built using the omixed framework, an architecture for supporting the rapid deployment of collaborative databases. More information about omixed and Terminizer, including a live demonstration of the service, is available on our website:

http://terminizer.org/

Employing ontologies in the annotation of data should not be unnecessarily burdensome to the user. Similarly, users cannot be expected to invest significant time in becoming intimately familiar with specific ontologies, many of which contain thousands of terms. To this end, we are investigating methods for assisting non-expert users in annotating their data.

We present a tool that automatically detects ontological terms in free text. Figure 1 illustrates the algorithm used by this tool. Once candidate terms have been identified the results are displayed either overlaid on the original text or in a list organised by ontology and frequency. Examples of these representations are shown in Figure 2.

The user can interactively accept or reject each match, or try to find a more appropriate match by exploring the network of ontology concepts themselves. In typical ontological resources, the parent(s) of a term represent broader concepts whilst the children of a term represent more specific concepts. In this way, the suggested match can used as a starting point for the user to find a more suitable term. Figure 3 illustrates the ontology browser interface, which has both a textual and a graphical mode.

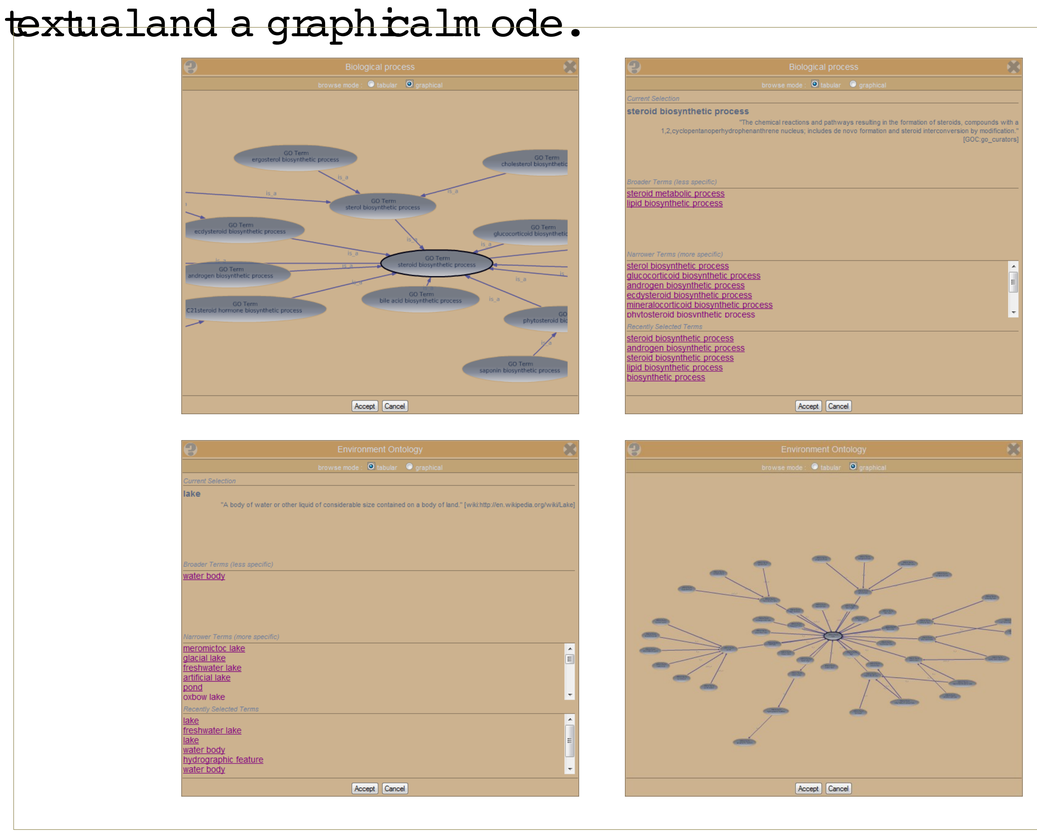

Figure 3 – The ontology explorers

Nature Precedings : doi:10.1038/npre.2009.3128.1 : Posted 22 Apr 2009