A workflow for developing biological ontologies using a document-centric approach

Aisha Blfgeh, Catharien Hilkens and Phillip Lord

School of Computing Science, Newcastle University

February, 2016

Domain specialists represent, manipulate and share their data in a wide-variety of different formalisms, and format. By far the most common, however, shared between almost every domain is free-text English and the office document. There is a significant gap between this and the tools used by ontology developers.

There are a number of different solutions to this problem. WebProtege, for example, presents a modified and customizable interface, which can be adapted to an individual community. Populous and Rightfield allow direct use of spreadsheets as a controlled vocabulary interface. Cellfie is a Protege-plugin which can transform a spreadsheet into an OWL ontology, which can then be developed further. These tools however much they support the use of office software, at some point, require leaving this software and moving into an ontology specific environment.

We are now investigating an document-centric workflow centred around the use of English and standard office tooling which we hope should ease the interaction between a domain specialist and an ontologist. In this case, we are investigating biomedical data. This form of data is not only heterogeneous but also require special knowledge to be dealt with. Therefore, ontologies are good for representing complex knowledge that is potentially changing and are widely used in biomedicine.

With this approach, biologists will read text, probably using Word, and input data into Excel, while ontologists will consume and generate these documents (using both manual and automatic techniques) into a computational representation. This transformation will be performed by Tawny-OWL – an environment for ontology development build using Clojure and the OWL APIs. Because it is highly-programmatic, we can make an arbitrary transformation, and additionally can perform this repeatedly, on-the-fly; this should enable us to maintain the original files as part of the ontology source code, while for Word files, we hope to consume these files, but then regenerate them as a form of ontology

documentation. We are currently testing these workflows using Tolergenic dendritic cell catalogue as a source in Excel Sheet format and docjure library to read and write Office documents in Clojure.
In order to evaluate and test the outcome of the system, domain specialists and ontologists will be used to utilise the system which comprise a feature of transformation a literate ontology forms into readable Word documents.