Ondex: data integration and visualisation for the Semantic Web

Catherine Canevet[1], Andrea Splendiani[1], Steve Kuo[1,3], Robert Stevens[2], Phil Lord[3], Chris Rawlings[1]

[1] Centre for Mathematical and Computational Biology, Rothamsted Research, Harpenden, UK
[2] School of Computer Science, University of Manchester, Manchester, UK
[3] School of Computing Science, University of Newcastle, Newcastle, UK

Many systems approaches to biology need to identify, integrate and analyse information that is captured in a myriad of databases which use a wide variety of different formats and access methods.

The Ondex data integration platform [1] (www.ondex.org) enables data from diverse biological data sets to be linked together, integrated, analysed and visualised using graph-based techniques. At the basis of Ondex is a graph data structure where entities and properties are associated to classes [2]. This data structure is closely related to the data model of RDF and supports a limited representation of ontologies.

In the context of the SABR project (http://www.ondex.org/sabr.html), we are investigating ways to create a mapping between the Ondex data structure and the RDF/OWL. Our objective is to allow Ondex to query, visualise and analyse Semantic Web based knowledge bases [3]. The challenges are to maintain acceptable performance when translating between the RDF and Ondex data models. Once this mapping is implemented, Ondex could also be used to build workflows for the curation and management of such knowledge bases.

The Ondex data model shares many similarities with RDF. For example, it has equivalents for Object Properties, Data Type properties, Types and Named graphs. On the other hand, one of the differences is that there are no global identities such as URIs available in Ondex. In fact, different information sources are aggregated by representing the original entities in “blank nodes” first and then by computing mappings among these nodes using their attributes.

We wish to devise a mapping strategy that will allow Semantic Web based knowledge bases to be exposed using the Ondex data model because of the rich feature set that Ondex supports. Among the features already implemented in Ondex that could be deployed on a generic Semantic Web based knowledge bases are:

- Support for the definition of workflows using an Ondex-specific engine, via Taverna or a scripting interface

- Methods to compute mappings between information resources, based on lexical information, unique identifiers, sequence similarity or text mining (mapping free text to entities and relations using information extraction).

- Methods to perform network analysis using betweenness and degree centrality measures (how influential an entity is in the network, how many relations it has) as well as a statistics module.

- Methods to interactively explore the content of integrated datasets as shown in Figure 1. A set of filters help users narrow down their integrated data sets to regions of interest in the network.

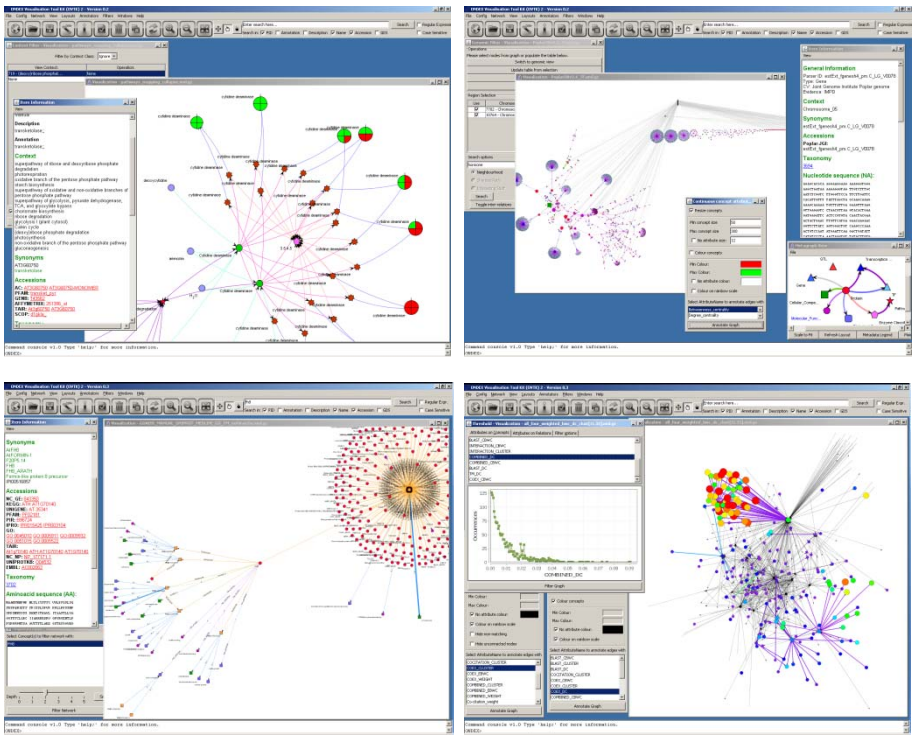

Fig. 1. Screenshots of the Ondex Visualisation ToolKit. Various filters and annotators allow users to visualise and analyse their datasets, integrated using Ondex's mapping methods.

Top-left: AraCyc, DRASTIC and microarray data. Pie charts show when genes are up/down regulated (red/green) given the selected treatments.

Top-right: Poplar genome sequence, PoplarCyc, UniProt, Pfam, GO, GOA and Medline. The genomic view allows user to select QTLs of interest and search their neighbourhood based on keywords and a given neighbourhood depth.

Bottom-left: GOA, UniProt, Medline and GO of *Arabidopsis thaliana*. Data integration and literature analysis methods are used to predict the function of previously unannotated genes.

Bottom-right: TAIR, IntAct and BioGRID. The resulting network shows the *Arabidopsis* interactome. The threshold filter displays the distribution of the degree centrality calculated for each protein and added to each protein as an attribute (previously by an annotator).

As a reference use case, we have started the investigation of how Ondex can be used to interact with BioGateway [4], which implements a Sparql endpoint. Apart from creating a mapping that would make BioGateway accessible in Ondex, we are defining a set of guidelines to harmonise the development of Ondex modules to ensure a common representation compliant with this mapping [5].

References

1. Köhler, J., Baumbach, J., Taubert, J., Specht, M., Skusa, A., Ruegg, A., Rawlings, C., Verrier, P., Philippi, S.: Graph-based analysis and visualization of experimental results with Ondex. Bioinformatics 22 (11):1383-1390 (2006).
2. Taubert, J., Sieren, K.P., Hindle, M., Hoekman, B., Winnenburg, R., Philippi, S., Rawlings, C., Köhler, J.: The OXL format for the exchange of integrated datasets. Journal of Integrative Bioinformatics 4, 62 (2007).
3. Antezana, E., Kuiper, M., Mironov, V.: Biological knowledge management: the emerging role of the Semantic Web technologies. Briefings in Bioinformatics 10, 392--407 (2009).
4. Antezana, E., Blondé, W., Egaña, M., Rutherford, A., Stevens, R., De Baets, B., Mironov, V., Kuiper, M.: BioGateway: a semantic systems biology tool for the life sciences. BMC Bioinformatics 10 (Suppl 10):S11 (2009).
5. Splendiani, A., Kuiper, M., Rawlings, C.: Toward a new paradigm for user interaction on the Semantic Web to support life sciences investigation. To appear in Proceedings of ISWC-SWUI (CEUR) (2009).