Linked Data services for Scientists exemplified using *Drosophila melanogaster* datasets

M.P. van Iersel, N. Anwar, L. Reynolds

General Bioinformatics, Reading, UK

1 Introduction

Semantic technologies allow scientists to rapidly integrate data from experimental datasets and online databases. On the computational side, the massive scale of these datasets poses some challenges. We find that there are additional challenges on the human side. Problems we encounter in practice include translating human questions to SPARQL queries, adjusting data models and ontologies to match the problem space of scientists, visualising data, and providing user-friendly access.

To address these issues, General Bioinformatics applies the following principles. Open source tools such as PathVisio and Cytoscape are adopted to provide visualisation options. Custom inferences layers are added to semantic data to address research questions. For example, we apply inferences on top of BioPAX pathways to enable the extraction of a bipartite biochemical reaction network with a single SPARQL query. Finally, we provide training materials for scientists for visualisation tools, ontologies, and SPARQL.

Our linked data system is supports life science researchers to answer research questions. For example, the system can gather supporting evidence from a variety of databases to identify interesting target proteins. Our public demo focuses on *Drosophila*, but the same principles apply to any model organism.

2 Results

We have integrated several *Drosophila melanogaster* datasets and built a public showcase. This showcase integrates diverse data types such as genes, proteins, FlyAtlas expression data and BioPAX pathways. This data is stored in a triple store running in a cloud instance. We provide several options for visualizing the data. First, a custom Cytoscape plugin can extract triple data and visualize it in network form. Secondly, a PathVisio plugin can be used to create visualisations on top of pathway diagrams derived from WikiPathways (See Fig. 1).

By following the demo, you should get a clear idea how linked data can be used to integrate diverse *Drosophila* datasets, how this can be visualized in pathway and network context, and how this can help to answer research questions.

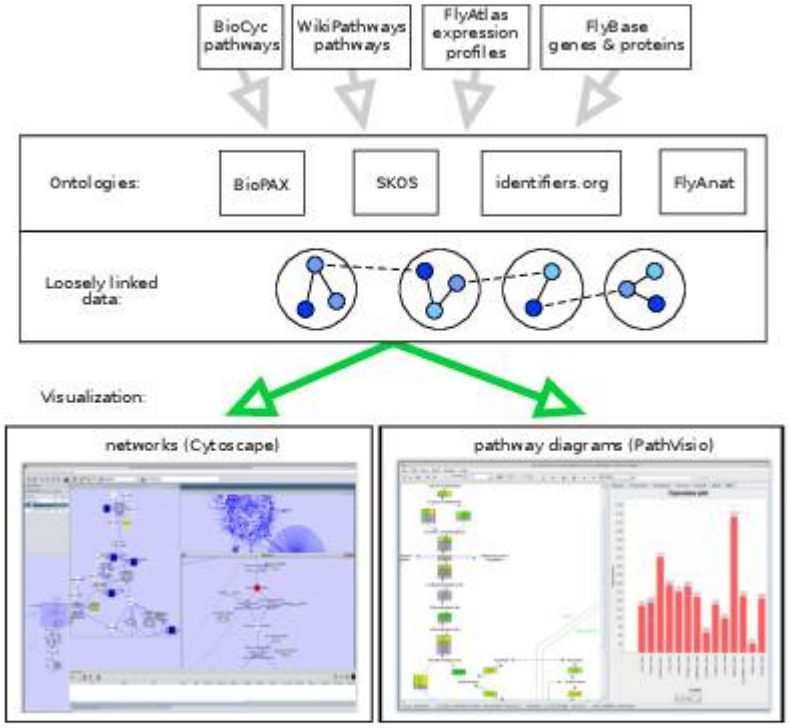

Fig. 1: a schematic overview of the integrated *Drosophila* data. A collection of databases (top) have been transformed into RDF and integrated using standard ontologies (middle). Two different visualisation tools are provided to create different views oft he data.

3 Availability

The integrated data, SPARQL tutorial and videos, RDF downloads, cytoscape session and live SPARQL endpoint are all available at http://fly.cloud.generalbioinformatics.com/