The State of the Nation in Data Integration

Carole Goble
Department of Computer Science
University of Manchester
Manchester, UK
+44 161 275 6195
carole.goble@manchester.ac.uk

ABSTRACT

Data integration has been discussed as a major challenge in bioinformatics for many years, ever since the beginning of dataset collection and distributed publication. A whole bunch of approaches have been tried over the years. We've had web-based systems that cross-linked data resources through massive cross indexing. We've had federated systems, ontology-based mediation systems, data warehouses, and workflows. We've seen lightweight protocols and heavyweight centralized systems; schema-reconcilation schemes and instance-linking through common vocabulary terms. We've seen data standards proliferate and web service standards promise a new future. Now we have data mashing and specialist wikipedia.

In 2002, Lincoln Stein argued that we were building a "Bioinformatics Nation" [1]. Here I'll survey the State of the Nation in Data Integration for Bioinformatics. I'll draw out common motifs for bioinformatics nationhood: from dictatorial systems to autonomous federated states to abject anarchy; closed vs open communities; /a priori/ planned economies /a posteriori /reunifications; revolutions, civil wars, strategic pacts, constituted unions and takeovers.

Throughout my talk I'll draw upon our extensive experiences of building data integration systems, from the TAMBIS, myGrid, myExperiment, ISPIDER, GIMS, e-Fungi, SeaLife, ComparaGrid and OntoGrid projects. I'll also refer to other influential data integration efforts like SRS, Integr8, DAS and caBIG. Many of these use semantic technologies in various ways. Other projects, like YeastHub are building RDF data warehouses. However, putting a bunch of RDF into a bucket is not the same as integration, just like exporting two databases as RDF doesn't necessarily mean you can link their content.

So I will ask the obvious questions. How does the Semantic Web help? How could it help? What needs to happen to help it help"? When doesn't it help? How do we enable Unity through Semantics?

REFERENCES

[1] Stein, L. Creating a bioinformatics nation. Nature Commentary 417, 119-120, 2002.