

NOVEMBER 11, 2006

## Weaving a Tapestry in eScience

[The Gaggle: An open-source software system for integrating bioinformatics software and data sources](#)

An interesting report on an OS app for integrating a variety of scientific databases and software programs.

Using the Gaggle we have integrated diverse databases (for example, KEGG, BioCyc, String) and software (for example, Cytoscape, DataMatrixViewer, R statistical environment, and TIGR Microarray Expression Viewer). This loose coupling of diverse software and databases enables simultaneous exploration of experimental data (mRNA and protein abundance, protein-protein and protein-DNA interactions), functional associations (operon, chromosomal proximity, phylogenetic pattern), metabolic pathways (KEGG) and Pubmed abstracts (STRING web resource). More importantly, the researcher can craft queries to explore these rich resources without any software constraints.

That's a mouthful. Even better is:

The Gaggle is a minimal, effective and open-ended system for integrating software and data sources used in systems biology analyses. The Gaggle's effectiveness comes from the recognition that four simple data messages each free of biological semantics, and a judicious use of the Java programming language, are all that is needed to integrate diverse types of data and software. More importantly, the Gaggle is easily extensible and new software and databases can be easily converted into geese of the Gaggle with little effort.

Now in language I can understand :-)

The Gaggle [5] uses a minimalist approach to integrate data and software. It is written in Java and uses standard Java libraries. It is simple to install, and easy to update; new data sources and software tools can be added with minimal implementation costs. A small server program (the 'Gaggle Boss') provides communication among analysis and display programs (the 'geese') which are modest adaptations of existing (or novel)

bioinformatics and computational biology programs and web resources. The Boss and the geese all run as separate programs on the user's desktop computer, communicating with each other, at the user's behest, by passing simple messages.

I have to admit that the details of this one are a little over my head, but the basic concept of tying disparate sources of data and application functionality in a simple communication framework is one my colleagues and I have been thinking a lot about in the last few months. It seems clear to me that the key to building effective eScience environments is doing just what this Gaggle of bioinformatics researchers have done: stitch together small best-of-breed modules into a cohesive whole, rather than build a monolithic application. But then that's what the new Web is all about ain't it?