

ISPIDER – A Pilot Grid for Integrative Proteomics



Birkbeck Project Leaders

Nigel Martin, Alex Poulouvassilis
[\[nigel.ap\]@dcs.bbk.ac.uk](mailto:nigel.ap@dcs.bbk.ac.uk)

Project Participants

N. Martin, A. Poulouvassilis,
 L. Zamboulis (Birkbeck)
 S. Hubbard, S. Oliver,
 S. Embury, N. Paton, C. Goble,
 R. Stevens, K. Belhajjame, J.
 Siepen (Univ. of Manchester)
 D. Jones, C. Orengo,
 M. Pentony (U.C.L.)
 R. Apweiler, H. Hermjakob,
 W. Zhu, C. Taylor, P. Jones,
 N. Vinod (E.B.I.)

Project Details

Funded by BBSRC.
 Duration: 3 years.

Project Web Site

<http://www.dcs.bbk.ac.uk/~lucas/projects/ispider>

Keywords

Bioinformatics
 Data Integration
 Grid Computing

Motivation

Integrated data analysis tasks across biological resources can provide a number of benefits, such as more reliable analyses by virtue of access to more data. It also relieves the biologist from having to have knowledge of each resource and reconcile their semantics and technologies.

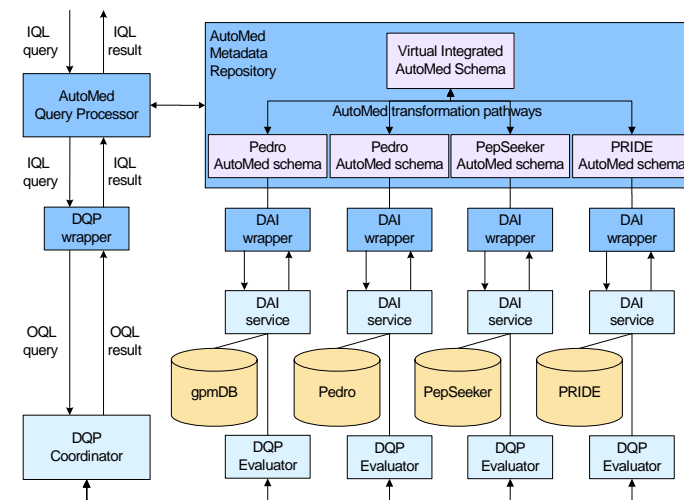
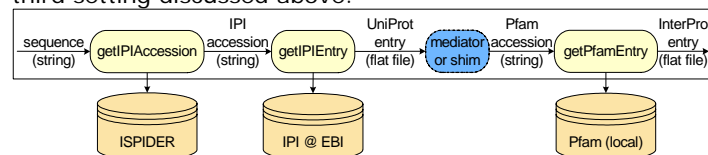
This objective requires solutions to the problems of *heterogeneous data integration* and *reconciliation of services* performing analyses over that data. Such services are created by parties worldwide using different representation formats and, as a result, semantically compatible services often cannot directly interoperate within a workflow. One of the outcomes of the ISPIDER project is a *uniform approach* to these problems that was applied the following settings:

- Virtual integration of relational resources in a Grid setting
- Materialised integration of (semi-)structured resources
- Reconciliation of services within a workflow tool

A Uniform Approach

Bioinformatics researchers frequently use workflow tools to perform complex experiments and analyses over local and remote datasets, such as local relational DBMSs and services producing flat/XML files.

In the figure below, a simple bioinformatics workflow presents two common workflow problems: (a) the first service needs to access multiple resources but the researcher is unfamiliar with their semantics, and (b) the output of the second service is semantically compatible with the input of the third one, but the two services cannot form a pipeline, due to representation format differences. The figures on the right illustrate the uniform approach applied in the first and third setting discussed above.



To address these problems, we use AutoMed to resolve the following heterogeneity issues, which are common to the problems of data integration and service reconciliation:

- Data model heterogeneity.

Different resources may use different data models, or one or more resources may not have an accompanying schema.

- **Semantic heterogeneity:** schema differences caused by the use of different terminology, or by describing the same information at different levels of granularity.

- **Schematic heterogeneity:** schema differences caused by modelling the same information in different ways.

- **Primitive data type heterogeneity:** caused by the use of different primitive data types for the same concept by different resources.

