DATA INTEGRATION AND INTEROPERABILITY

ETI BioInformatics

Solving Problems when Connecting Databases to Distributed Systems

Authors: ir. W. Addink and drs. M.L. Brugman

Distributed systems in biodiversity

At present, distributed systems in biodiversity consist of loosely coupled fully autonomous databases on which only distributed search and retrieval of data is possible. This is called distributed searching – where a user sends a query to multiple sites at once, and the owners of the data return results from their own data stores. These systems may lead to distributed database systems: a set of databases stored on multiple computers that typically appears to applications as a single database. In such a system, an application can simultaneously access and modify the data in several databases in a network.

Examples

Several architectures for distributed systems in biodiversity are already implemented or in development like:

- The GBIF network,
- BioCASE,
- MaNIS,
- OBIS,
- BioMOBY.

Autonomy of data sources and data receivers

These systems show a lot of similarities but also differences. All current systems consider the autonomy of the data sources as vitally important, however the autonomy of data receivers (end users) is taken less into account. Data receivers are autonomous in the sense that they typically have different information needs and vary in the ways they perceive their particular domain of interest. For them a “one integrated schema fits all” is not satisfactory [1]. Also, end users often do have little need for ‘raw data’ like specimen or observations but instead need interpretations from calculated results based on this raw data.

Data heterogeneity

Most current networks use a common concept schema like ABCD or Darwin Core. Less often implemented are standardisation of data itself, and things like duplicate elimination and missing value substitution. This leaves some data heterogeneity existing and therefore some interoperability problems still exist in the current systems. With the currently implemented distributed systems in biodiversity, development is moving into the direction of webservices.

Webservices for interoperability

Webservices seems to be one of the best solutions at current to achieve interoperability. Being application-centric, webservices are:

- more scalable then a human-centric webportal approach,
- can be very language and system independent,
- easier to establish than for instance a GRID network.

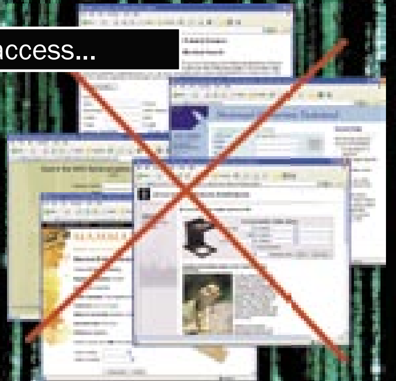
Webservices have the promise to provide a new level of interoperability. Its interoperability goals are to provide seamless and automatic connections from one application to another. SOAP, WSDL, and UDDI protocols define a self-describing way to discover and call a method in a software application – regardless of location or platform. However developers have to interpret the meaning in parts of the specifications which still allows interoperability problems to creep in at the discovery, definition and request/response mechanisms [2]. UDDI and WSDL also seem to be not very useful in getting a machine to know which services it can use.

References:

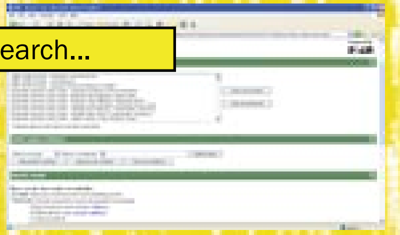
[1] Patrick Ziegler. USER-Specific Semantic Integration of Heterogeneous Data: What Remains to be Done? Technical Report ifi-2004.01.

[2] Frank Cohen. Understanding Web service interoperability. Issuses in integrating multiple vendor Web services implementations, February 2002.

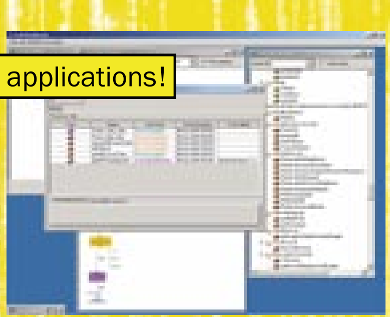
DATA INTEGRATION AND INTEROPERABILITY



from individual access...



...to distributed search...



...even directly in applications!