# FEDERATED INFORMATION SYSTEMS FOR COMMUNITIES

David J. Abel
*CSIRO Mathematical and Information Sciences*
*GPO Box 664, Canberra, ACT 2601, Australia*
Dave.abel@csiro.au

## 1. EXTENDED ABSTRACT

One of the more exciting opportunities enabled by the Internet is sharing of data and software within communities of interest. The early waves of exploitation have provided faster and easier communication, through email and through Web publishing of research results, research materials (including data and software), news and general information. For individual researchers, this has enabled access to a richer set of resources and fostered cooperation within specialist groups. A strengthening trend is for individuals or institutions to share resources (both data and software) through Web services. To search for available information on a gene, for example, a biotechnology researcher can now access a remote genomic database, rather than incur the costs of maintaining her own local database. This paper envisages the establishment of collections of Web services, by a community of interest and for the community of interest. Technically, this form of Federated Information System poses the usual problems of size (the number of services present) and heterogeneity in the services. Although recent advances in the area give some confidence that large-scale, highly-heterogeneous networks of Web services are within reach, the social aspects of establishing shared resources require reconsideration of some

X. Zhou et al. (eds.), *Visual and Multimedia Information Management*

fundamental elements of architectural design. This paper examines some informations systems research issues in design of federated information systems for large communities of interest.

In this paper, we will use as case studies three representative communities of interest. We will use our experience in design and implementation to observe and generalize some common elements that are influential in the design of information systems. These communities are an association of local government authorities in Sydney , the astronomical research community, and a consortium of Australian human health research institutions. Information systems developed or under development are the Sydney Information Highway, the Virtual Observatory and the Australian Health Research Network, respectively. These consortia are essentially defined by common goals and an acceptance of benefits, to the members and to their stakeholders, from a collective effort.

These common goals are generally loose, rather than able to be defined in terms of a crisp set of objectives or common processes. For example, an astronomer will develop greater insight into a stellar object by fusing data from many sources. Fusing the data will typically require intelligent selection of data sources, analysis and visualization tools, rather than following a well-defined process. While some standard processes can be established, often the network acts as a resource collection, enabling many processes, many of which are combined on a once-off basis.

The success of the community's efforts is vitally dependant on achieving a critical mass of resources within a reasonable time. The extent and vigour of participation by an individual member depends on the benefits to the member, peer-group and stakeholder expectations on contributions, continuing competition between members, and the costs of participation. This leads to a social structure of the community as a loose federation, with weak centralised coordination and management. This contrasts a social structure in which a central unit is mandated to establish the resource (including metadata), perform integration and establish standards.

Together these factors suggest that the well-known methodologies for establishing the resource collection and achieving interoperability have limited applicability. Approaches based on a global schema, process models and comprehensive standards are all highly problematical. Three clusters of issues are particularly interesting.

Firstly, there is the issue of describing services in a highly-heterogeneous collection. The importance of minimizing the barriers to participation leads to the desirability of allowing a member to contribute a service 'as is, where is', rather than extending existing capabilities or migrating to a new, standards-compliant environment. It is also intuitively undesirable to inhibit a member from offering capabilities that are special in some way. The technical challenge is then to accommodate differing capabilities in data sources, in terms of selection operators, both in description of the services and in any command language. A further open question is in the modelling of applications, such as statistical routines in the Australian Health Research Network.

Secondly, data integration across services has some complexities. In our case studies, it appears infeasible to capture a global schema, because it would be very large indeed and because the set of services will be very dynamic. Rather it is more likely that the activities of the members of the community will lead to publication of collections of purpose-specific views materialising sets of objects of interest. In some cases, linkages across data sources will be facilitated by common keys for objects (such as the astronomical designations of the International Astronomical Union), or by standardized representations of values in some domains (such as the coordinate systems of geospatial information systems). In other cases, linkages will need to be by joins, using both the common relational operators and externally-defined join relations. Tools to discover views and means of publishing them in usable ways will be particularly valuable. The local-as-view schema definition approach will be an important element.

Finally, there is the question of effective use of the large-scale resource that could be assembled collectively by the community. In addition to the data sources, there could be software and high-performance computing systems. Resource discovery, task planning (including optimization) and task execution are topics likely to reward re-consideration of the current techniques.

## BIOGRAPHY

Dave Abel has contributed to the fields of spatial access methods, spatial databases and multidatabase systems. He is currently Chief Scientist in CSIRO. His current work centres on Web services.