I2T: An Information Integration Testbed for Digital Government

PI: Chaitan Baru
Co-PI’s: Amarnath Gupta, Ilya Zaslavsky, Yannis Papakonstantinou (UC San Diego), Peter Joftis (U. Michigan)

The broad goal of the I2T project was to develop XML-based data mediation technology for integrating geospatial and statistical data, which is of great interest to a number of statistical and other government agencies. The research participants included the San Diego Supercomputer Center and Computer Science Department at the University of California, San Diego, and the Inter-University Consortium for Political and Social Science Research (ICPSR) at the University of Michigan. The initial agencies that were identified as potential beneficiaries of this research included the Bureau of Labor Statistics (BLS) and Census Bureau. The role of agencies was identified as providing not only the application drivers but also advice and feedback. Interactions with agencies were primarily via meetings held at in Washington or at San Diego Supercomputer Center. During the course of this project, there were also interactions and collaborations with other NSF Digital Government projects such as the Iowa State project on field data collection (Sara Nusser, PI) and data integration and the University of Maine project on spatial metadata (Peggy Agouris, PI). Indeed, these collaborations also eventually led to a funded joint NSF ITR project on geospatial data grids (GeoGrid), involving SDSC, Iowa State and Maine.

The research goals of this project were to develop wrapper/mediator-based systems for Web-based integration of distributed databases, and to develop innovative query interfaces for statistical and GIS data integration. Some of the outcomes of this research include development of XML-based query language and query processing techniques including, XMAS (XML Matching And Structuring language), which was in fact introduced prior to the advent of the Xquery standard; introduction of the DOM-VXD (Virtual XML Document) concept to facilitate incremental query evaluation in support of a query cum browsing model of interaction; development of an innovative, SVG-based client-side mapping software; development of techniques for integration of map-based (GIS) data from Internet map servers using a combination of XML mediation and Web-based workflows; and, development of Web services to access some of the ICPSR datasets using the XML-base Data Documentation Initiative (DDI).

The project was organized as a collaborative effort among UCSD/SDSC and U. Michigan. Points of contact were identified at the Census Bureau and BLS. Towards the later part of the project we also established contact with the EPA. Primary communication between the project and agency partners was via meetings, sometimes as a group and sometimes among individual on both sides. Communication within the project was via email and annual meetings at SDSC. One of the collaborations between I2T and the Census Bureau was with an internal Census project at that time on metadata-based search and access to data sets. The I2T approach, using XML-based querying of data/metadata, and integrated access to information was relevant to this effort. This led to a separate NSF SGER grant to evaluate software architecture being developed by Census and provide technology
insertion, where possible. Also, the collaboration between SDSC and U.Michigan/ICPSR led to technology transfer from SDSC to ICPSR and transfer of knowledge about DDI from ICPSR to SDSC.

A related technology spinoff from this project was Enosys Software Systems co-founded by co-PI Yannis Papakonstantinou. Some of the graduate students funded by the I2T project became key technical members of Enosys. XMAS and DOM-VXD were adopted as technologies in the Enosys Xmediator product. Some of the Enosys technology was eventually acquired by BEA as part of their Web Logic Liquid Data offering. Another software product was AxioMAP, developed by Senior Personnel Ilya Zaslavsky, which is an SVG-based, client-side mapping tool. This software has been adopted and used in several other projects at SDSC and elsewhere, including a Digital Government project at U.Wisconsin. The SVG technology is also of interest to ESRI, a GIS software company and a research partner at SDSC. A demonstration of this capability in conjunction with ESRI products (ArcIMS) is currently under way.

The information integration challenge being addressed by I2T attracted the interest of other technology partners at SDSC who also joined this effort. This includes Polexis, a provider of Java-based, application-level integration software, who participated in an I2T technology demo at the Digital Government 2002 Conference. Another was BlueTitan, a provider of Web services management software. Their software, BT Studio, was used to develop Web services over GIS information sources and provide workflow-based information integration using these Web services. The GIS information integration technologies developed in I2T have been further enhanced and are now being used in other research projects, e.g. GEON (The Geosciences Network), which is an NSF large ITR project.

One of the challenges we faced in the project was the inability to successfully deal with the distance barrier—between San Diego and Washington DC. As a result, there was patchy communication between project researchers and agency partners/advocates. Also, agency partners were sometimes busy with major, new internal IT efforts and it was not easy to put the I2T research project high on their priority list, since the goals of I2T were more long term (e.g., 3 years as opposed to 3-6 months).

The research value in working with an application-oriented program like Digital Government is the exposure to “real” problems and real concerns. While many of the requirements in “real world” applications could be met with IT “best practices”, close collaboration with DG partners also exposes those areas that need true innovation and CS/IT research. This helps CS/IT researchers focus their work on more realistic problems.

To further improve the Digital Government program, it is important for agencies to realize that the people who serve as the interfaces or points of contact, need to have some release time from regular duties to focus on the project so they have the time to think more broadly of solution approaches. There should be explicit outreach from each project to a broader range of folks at the agency other than just the individuals who happen to be participating in the project.