Databases in Virtual Organizations: A Collective Interview and Call for Researchers

Marianne Winslett
Database and Information Systems Laboratory
University of Illinois at Urbana-Champaign

As an untenured faculty member, I can't afford to try to convince the database research community that a problem is important. A great outcome of the DIVO workshop would be to convince the community that this problem is important. *–Anonymous workshop attendee*

When the Databases in Virtual Organizations (DIVO) workshop convened after SIGMOD 2004 in Paris, many of us attending weren't sure what a virtual organization was, much less what relevance it could have to database research. Five hours later, as the lights snapped off in the rest of the building and the maintenance crew hovered patiently outside our meeting room, we had become a group with a mission: to let the database research community know what an incredible idea generator and testbed virtual organizations could be for research on information integration and data security.

A virtual organization is a set of collaborating organizations working toward a common goal. The collaboration may last a short or a long time, and has no centralized control. The members and activities of the collaboration evolve dynamically, often rapidly. The main force behind the development of virtual organizations comes from business, which is facing intense competitive pressures in the global economy. These pressures push businesses to decouple business needs from the means of satisfying those needs, by developing the ability to quickly determine the best way of meeting a particular internal business function, and just as quickly to switch from their current way of meeting the need to a new approach that is better---as quickly as a click of a mouse button. As a small example, an organization might like to continuously and automatically monitor the price and performance of its chosen express mail carrier and the carrier's competitors, and switch to another carrier if that carrier will offer better price-performance. Businesses are under pressure to reorganize as virtual enterprises, by adopting a model where every business need, even those currently being satisfied internally, could undergo this continuous evaluation and potential reassignment. The shift toward virtually organized enterprises has been enabled by recent advances in technology, including the rise of the Internet, automation of many routine business processes, the development of standard interfaces for those processes, the trend toward shifting corporate alliances, and the ability to relocate facilities easily.

If you are not already familiar with virtual organizations, you may be surprised to hear that they already have killer apps: supply chain management, enterprise resource planning, and customer relationship management. For example, www.businessweek.com says that the adoption of supply chain management will raise the earnings of a $100 million dollar company by up to 6% a year---an irresistible carrot for the company. Supply chain management and the other killer apps all require information to flow across (sub)organizational boundaries, a hallmark of virtually organized enterprises. For example, when taken to its logical conclusion, companies can use supply chain management to look into the databases of their suppliers' suppliers, and their customers' customers.

Information and process integration are key technology issues in virtual organizations.

Raghu Ramakrishnan:

"Virtual organizations exist, but they aren't as virtual as you might think.

"Even in dynamically established virtual organizations, you must have negotiated agreements and workflows that enforce them and are set up long in advance."

Rakesh Agrawal:

"In the system we implemented [for Hippocratic databases for virtual organizations], there is a negotiation phase up front where people agree that a certain amount of info can be shared with one another."

The database research community is already well aware of the importance of information integration. However, information integration is just one piece of the picture---the larger issue is business process and task integration, of which information integration is just one component. Businesses need to integrate all the components of an entire task: messages, business processes, workflows, policies, and data. This is not a new area of endeavor; what is new is the speed at which integration has to be accomplished, by non-expert integrators who really need automated assistance. Already integration problems represent over half of a typical IT budget in a Fortune 500 or government organization. The total amount of data in business process objects is huge (e.g., 15,000 relational tables used to represent only 150 business objects).

Workshop participants concluded that virtual organizations raise no new issues in information integration; instead, they exacerbate current known problems and thus point clearly to future research directions. In sum:

+ Due to their dynamic nature, virtual organizations make on-the-fly data integration extremely important. No one will have six months to set up a full-blown data integration system---organization members will need their answers quickly.
+ Those who integrate data in virtual organizations will not be experts and will only do "best effort" work. Thus approximate integration efforts (e.g., semantic mappings that are roughly correct) will be crucial.
+ In virtual organizations, integrating data is important, but so is integrating business processes, policies, messages, etc.

Hamid Pirahesh:

"Business process integration is vital [for modern businesses]. BPEL is a standard for helping with this, and it was done without [the database research community's] participation. **We are not participating in determining the schema, constraints, behavior of industry standard business objects.** There are no formal semantics for these business objects. It's an elephant now, and we needed to be there at the beginning and provide a better theoretical foundation.

"The data in the database is just the assembly language level of business objects. This is why 'objectification' is so important (providing a semantic model closer to what business users care about).

"ETL [Extraction, Translation, and Load] and analysis will be very important. As Stonebraker said [in his keynote speech at SIGMOD 2004], you can work in this area only if you have tenure. So you don't see much ETL research presented at SIGMOD.

"With virtual organizations, we must expect that data leaks out to the other side of the world, where it is subject to a different set of laws. How do you deal with the access control and security? How do you deal with various versions of schemas at the same time? When the schema of a business object evolves, you cannot force the whole world to update their schemas simultaneously. So, you have to deal with different versions of objects. You have to deal with new and unexpected types of objects flowing into your system every day. You have to discover the schema of objects by pattern matching and correlation with reference data and what you already know. I call this `schema chaos,' and we must thrive on chaos. Unlike traditional information systems, there is no single DBA in control any more."

Phil Bernstein:

"Scalability [of virtual organizations and their associated information integration] is a big issue. On 9/12 they had the pictures of the [World Trade Center airplane] hijackers on the front page of the newspapers. I wonder how long it would have taken [to find those pictures] if even one more database was involved.

"We need tools to establish bindings between applications, databases, and processes, and tools to evaluate them so that we can discover errors along the way (e.g., as schemas evolve, or during a crisis when you find you are connected to the wrong database). This requires agreement up front to come up with the bindings, and an investment in wrapping certain existing important applications so that they connect to others.

"The competence of the designers [of virtual organization information systems] will not be expert-level. [Their work] needs to be a "best effort" effort; there will be errors and they must be tolerated, and fixed afterwards.

"**If we worked on some real application problems, I think we might discover interesting problems that would have higher impact than what we might think of a priori.** [For example,] GIS systems are often produced by separate vendors, and they are integrated by a human clicking on separate screens.

System building is the entrée to this research area.

Raghu Ramakrishnan:

I agree with Phil on everything. <laughter>

Catriel Beeri:

"The growth in memory is not so important as growth in communications speed. We need to learn how to share *abstractions*--- that's the big issue."

The flow of information across organizational boundaries raises new security issues. How can Walmart and Widgetcorp describe their own information dissemination policies? Will Widgetcorp understand Walmart's policies, and vice versa? How can we efficiently ensure that the policies stay with the data wherever the data goes, and how can we audit policy compliance? What happens if the policies themselves are sensitive business assets? How can a user from Widgetcorp prove that she satisfies one of Walmart's policies?

Beyond identifying particular areas of research that are of importance in virtual organizations, DIVO participants have strong opinions about the best way to go about that research. In the integration area, we recommend a bottom-up approach: choose a real-world example of appropriate size, solve it, and then generalize from your solution. If you are able to get access to supply chain management, enterprise management data, so much the better: these applications are widely used, they are of key economic importance, and they offer great possibilities for testing ideas in data security and information/process integration, as well as for generating such ideas. The bottom-up approach also addresses the widely-expressed sentiment of workshop participants that we have many of the components needed to solve virtual organization problems, but we don't know how to put them together.

Your friendly local Fortune 500 company may not be willing to share its supply chain management data with you, but you are likely to find willing cooperation on a smaller scale from your local government, educational, or charitable organizations. These organizations may not be running supply chain management software, but they will be facing virtual organization information integration problems. For example, the police and fire departments in Champaign, Illinois, would love to be able to get live feeds of sensor, video, and other data from locations where an emergency is in progress. In some cases, the city owns the data sources, but more often, the sources belong to a separate organization, and outside access raises security and on-the-fly information integration concerns.

Bhavani Thuraisingham:

"I would like to see ideas about virtual organization security in proposals sent to NSF.

"We have the components [needed to create virtual organizations], but we don't know how to put them together, [and we are not entirely sure that they are the *right* components].

"Colleagues at NSF say that we reapply the same techniques to each new area: RDBMS, OODBMS, XML, etc.

"The danger is for us to get all unfocused because there are so many different research directions on this list. We should not just start another research area."

DIVO participants postulated that the integration activities of virtual organizations should take place in the context of an agreed-upon backbone of technology---a **collaboration bus**. The collaboration bus is an a priori infrastructure that will allow users to quickly plug in new collaborators, and allows them to broadcast and advertise their identities, capabilities, policies, constraints, workflows, demands, and requests. The collaboration bus is a place where matchmaking can take place through private and robust communication channels, relying on trust mechanisms and policy enforcement.

Research activities in peer-to-peer computing, the semantic web, grid computing, and web services are all relevant to virtual organizations. However, DIVO participants characterized some of that research as "paperware," and recommended that the database community not rely on these other areas to provide solutions to all the needs of virtual organizations.

A recurring theme in the DIVO discussions was that we have many components that might be used to address the problems of virtual organizations, but we don't know how to put those components together, and we aren't even sure that they are the right components. We felt that the way to address that issue is to do some application-driven system-building. At the same time, we noted the lack of representation of large systems projects in the SIGMOD proceedings: the days of Ingres, Postgres, System R, and Shore are over, at least for untenured faculty members.

As may be gathered by the above remarks on research directions and methodology, the half- day DIVO workshop devoted a large amount of time to discussions. With 20 attendees, the group size greatly facilitated give and take during discussion periods and during the closing panel/audience discussion. In addition, we kicked off the workshop with several talks of a tutorial nature. The workshop organizers presented introductions to virtual organizations and to crisis response, and Cyrus Shahabi gave an excellent and visually compelling overview of the issues in integrating geospatial data with other information during crisis response. We also had an interesting session of contributed papers, including presentations on sovereign information sharing from IBM Almaden; mass collaboration in information sharing from UIUC; and automated service integration during crises, from the University of Pittsburgh. The complete proceedings, including the tutorial slides, can be found at http://dais.cs.uiuc.edu/divo2004/ . The extensive notes that we took during the discussion sessions were later coalesced into this document. As editor of the notes, I could not help noticing how many compelling remarks were made during the discussions, and I have sprinkled them in sidebars throughout this narrative, with attribution. In closing, I would like to thank all the workshop participants, for such great discussions; my co-organizers Sharad Mehrotra and Ramesh Jain, for their work in putting the workshop together; the workshop presenters, for their thought-provoking ideas; Alon Halevy for his many thought-provoking questions during the first half of the workshop; and SIGMOD, for sponsoring the workshop.

Workshop closing dialogue: where do we go from here?

Sharad Mehrotra: Would it be helpful for the DB community to push the notion of a virtual organization?

Raghu Ramakrishnan: We need to create compelling examples. Look at the companies, see what they sell, see where they fall short; that is the gap that we should fill, e.g., in supply chains.

Rakesh Agrawal: You won't *find* examples because the facilitators aren't there. It is a chicken and egg situation.

Phil Bernstein: **We [workshop participants] need to lead by *doing*, by trying it ourselves.**