

Data-driven Office Dynamics

Emerging field of Information Visualization helps archivists make sense of organizational records

As we move toward a paperless world, professional archivists are experiencing a revolution. Once the keepers of rare paper documents, they are now faced with massive digital proliferation and questions of how to manage vast streams of data.

According to Maria Esteva, a PhD candidate at The University of Texas at Austin School of Information, this new era offers both challenges and opportunities for archivists (also known as information scientists).

- Information Visualization (or InfoVis) helps archivists understand complex digital records using high-performance computing resources.
- TACC collaborated with Maria Esteva to interpret an unstructured archive of organizational documents that shed light on changing office dynamics.

“The quantity of paper records and the question of where to store them was a big reason why archivists became selectors of records, because there was no room,” Esteva said. “Now, potentially, we can store all these electronic records, so our role has evolved into determining what these records can tell us about the people and the organizations that created them, and devising approaches to make sense of them and make them accessible to people.”

Esteva chose a challenging archive for her research subject: 17,000 documents, in Spanish, accumulated over ten years, charting the tenure of an important Argentine philanthropic organization where she once worked. Though the task of interpreting an unstructured electronic text archive in Spanish appeared daunting, Esteva was confident she could use the power of text mining to uncover the hidden relationships among the records and through them the relationships between their creators.

Generally, the bulk of paper records in an organization, such as legal documents or official correspondence, are kept in a structured record-keeping system. And while such records also exist in paper version in the Argentinean archive, Esteva was interested in the wealth of unstructured information present in the shared directory on the organization’s networked server: drafts, fragments, reports, communications, and even personal employee records, that could shed light on the work of the organization.

Though the task of interpreting an unstructured electronic text archive in Spanish appeared daunting, Esteva was confident she could use the power of text mining to uncover the hidden relationships among the records and through them the relationships between their creators.

Because the volume and nature of unstructured digital documents can overwhelm traditional archival methods, Esteva applied automated computational tools to sort the records by employee and year and text-mined ten yearly sets from 1996 to 2005.

“Through text mining, the documents in each set are transformed into ‘a bag-of-words representation,’” Esteva explained. “It keeps all the statistics, telling you how many times each word appears in each record.” The bag of words representation calculates the similarity between

every record and averages the similarities between paired employees to measure the strength of the relationships between them, based on the records that they wrote and co-wrote. The method helped Esteva obtain vast numerical matrices, but without a dynamic representation of the data, she found it difficult to interpret her results.

After learning about the visualization services at the Texas Advanced Computing Center (TACC), Esteva decided to see if TACC's experts could help her understand and communicate her data. "TACC has the skills that I do not have, and I have the data and the knowledge about this projects, so we complement each other," Esteva said.

“When I was doing my visualizations in a static form, I was really very limited by the software and what it was telling me I could do,” Esteva recalled. “Working with TACC liberated me from that. Now I’m not constrained. We find solutions.”

Esteva used *Maverick* (a terascale remote visualization system, consisting of a Sun E25K with 128 processors) to turn her complex datasets into easy-to-interpret movies. *Maverick* is one of the key remote and collaborative visualization systems on the TeraGrid, the open scientific discovery infrastructure combining leadership class resources at eleven partner sites to create an integrated, persistent computational resource.

Esteva's analysis converts the relationship between documents into relationships among people, and shows how the office dynamics changed over the course of a decade. The movies that she developed with the assistance of Paul Navratil, one of TACC's visualization specialists, focus on the evolving relationships within the organization and the interactions among co-workers and departments.

“What we see over ten years are the dates that people entered and left the organization, what areas and employees worked closer and when, who remained in a consistent relationship over the years, and what happened during the last years as the organization was approaching its closure,” Esteva enumerated.

After checking her conclusions with the employees of the organization, she determined that her results agreed with their experiences. The visualizations show changes in the office's dynamics beyond the official organizational chart and reflect an ever-changing organization.



Created by Esteva and Navratil, this visualization plots the relationship between the director and the employees of an organization relative to the strength of their working relationship. To watch the animation, click on the image.

“When I was doing my visualizations in a static form, I was really very limited by the software and what it was telling me I could do,” Esteva recalled. “Working with TACC liberated me from that. Now I’m not constrained. We find solutions.”

Navratil’s work with Esteva is one of TACC’s first forays into the emerging field of Information Visualization (InfoVis), where information — rather than scientific data — is analyzed and represented visually. The project helps Esteva find answers to her research topic, and also expands TACC’s expertise to better help future users.

“Not only can Maria get the conclusions she needs out of her database, but we at TACC can create software that we can then apply to other scientists’ data and make open to the scientific community to multiply the impact,” Navratil said. “We don’t want to create a one-off visualization. We want to make something that will continue to reap benefits not just for us and Maria, but for other scientists as well.”

Recently, Esteva and Navratil began exploring whether they can use the methods and tools they’d created to probe the Enron email archive (a corpus of 500,000 emails from 150 Enron employees) and visualize the working relationships that preceded the company’s collapse.

“Once we build this framework for Maria’s data, we can then apply it to the Enron email archive and see if we can generate a similar organizational visualization that shows something interesting about how the key players in the Enron system came together and went apart,” Navratil said. “That would be a great validation of our work and would allow us to assess the scalability of the text mining methodology.”

The value of this research lies not only in its ability to show connections through documents, but also in its inclusion of a large proportion of digital materials that are at risk of being overlooked under the current paradigm of regulated recordkeeping. By including unstructured, as well as structured, archives, one can better reflect the role played by everybody in the organization, from the president to the receptionists.

“We need ways to understand the digital imprint we’re leaving,” Navratil said. “We have these computers that already can read and manipulate digital data, so now it becomes the challenge of information and visualization scientists to map the computer data and put it back in human terms.”

This work has been showcased in [“New Skills for a Digital Era”](#) (Available from the Society of American Archivists (SAA)), [“Texts and Bitstreams: Appraisal and Preservation of a Natural Electronic Archive”](#) and [Digital Humanities 2007](#).

Aaron Dubrow
Texas Advanced Computing Center
Science and Technology Writer
April 23, 2008