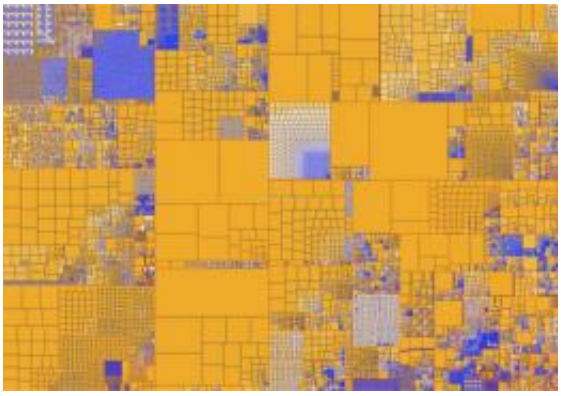
A glimpse of the archives of the future

April 6 2011, By Aaron Dubrow



Presentation of the entire testbed collection represented as a treemap in which the archivist can assess correspondence between number of files (size of the directory) and size of files (ranges of yellow) and their distribution in directories. Credit: Visualizations courtesy of Maria Esteva, Weijia Xu, Suyog Dutt Jain, and Varun Jain.

How does an archivist understand the relationship among billions of documents or search for a single record in a sea of data? With the proliferation of digital records, the task of the archivist has grown more complex. This problem is especially acute for the National Archives and Records Administration (NARA), the government agency responsible for managing and preserving the nation's historical records.

At the end of President George W. Bush's administration in 2009, NARA received roughly 35 times the amount of data as previously received from the administration of President Bill Clinton, which itself

was many times that of the previous administration. With the federal government increasingly using social media, cloud computing and other technologies to contribute to open government, this trend is not likely to decline. By 2014, NARA is expecting to accumulate more than 35 petabytes (quadrillions of bytes) of data in the form of electronic records .

"The National Archives is a unique national institution that responds to requirements for preservation, access and the continued use of government records," said Robert Chadduck, acting director for the National Archives Center for Advanced Systems and Technologies.

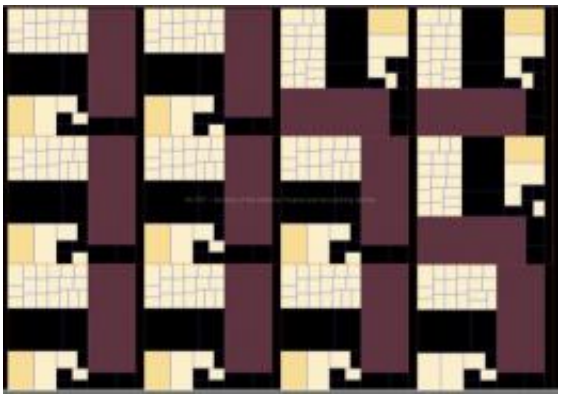
To find innovative and scalable solutions to large-scale electronic records collections, Chadduck turned to the Texas Advanced Computing Center (TACC), a National Science Foundation- (NSF) funded center for advanced computing research, to draw on the expertise of TACC's digital archivist, Maria Esteva, and data analysis expert, Weijia Xu.

"For the government and the nation to effectively respond to all of the requirements that are associated with very large digital record collections, some candidate approaches and tools are needed, which are embodied in the class of cyberinfrastructure that is currently under development at TACC," Chadduck said.

After consulting with NARA about its needs, members of TACC's Data and Information Analysis group developed a multi-pronged approach that combines different data analysis methods into a visualization framework. The visualizations act as a bridge between the archivist and the data by interactively rendering information as shapes and colors to facilitate an understanding of the archive's structure and content.

Archivists spend a significant amount of time determining the organization, contents and characteristics of collections so they can

describe them for public access purposes. "This process involves a set of standard practices and years of experience from the archivist side," said Xu. "To accomplish this task in large-scale digital collections, we are developing technologies that combine computing power with domain expertise."



This snapshot corresponds to a regularly organized website containing a total of 2,000 files of different file formats. Highlighted in shades of yellow are different number of Portable Document Format (PDF) files. The purple color shows patterns in file naming convention across directories. Credit: Visualizations courtesy of Maria Esteva, Weijia Xu, Suyog Dutt Jain, and Varun Jain.

Knowing that human visual perception is a powerful information processing system, TACC researchers expanded on methods that take advantage of this innate skill. In particular, they adapted the well-known treemap visualization, which is traditionally used to represent file structures, to render additional information dimensions, such as technical metadata, file format correlations and preservation risk-levels. This information is determined by data driven analysis methods on the visualization's back-end. The renderings are tailored to suit the archivist's need to compare and contrast different groups of electronic

records on the fly. In this way, the archivist can assess, validate or question the results and run other analyses.

One of the back-end analysis methods developed by the team combines string alignment algorithms with Natural Language Processing methods, two techniques drawn from biology. Applied to directory labels and file naming conventions, the method helps archivists infer whether a group of records is organized by similar names, by date, by geographical location, in sequential order, or by a combination of any of those categories.

Another analysis method under development computes paragraph-to-paragraph similarity and uses clustering methods to automatically discover "stories" from large collections of email messages. These stories, made by messages that refer to the same activity or transaction, may then become the points of access to large collections that cannot be explored manually.

To analyze terabyte-level data, the researchers distribute data and computational tasks across multiple computing nodes on TACC's high performance computing resource, Longhorn, a data analysis and visualization cluster funded by NSF. This accelerates computing tasks that would otherwise take a much longer time on standard workstations.

"TACC's nationally recognized, HPC supercomputers constitute wonderful national investments," said Chadduck. "The understanding of how such systems can be effective is at the core of our collaboration with TACC."

The question remains as to whether archivists and the public will adapt to the abstract data representations proposed by TACC.

"A fundamental aspect of our research involves determining if the

representation and the data abstractions are meaningful to archivists conducting analysis, if they allow them to have a clear and thorough understanding of the collection," said Esteva.

Throughout the research process, the TACC team has sought feedback from archivists and information specialists on the University of Texas at Austin campus, and in the Austin community.

"The research addresses many of the problems associated with comprehending the preservation complexities of large and varied digital collections," said Jennifer Lee, a librarian at the University of Texas at Austin. "The ability to assess varied characteristics and to compare selected file attributes across a vast collection is a breakthrough."

The NARA/TACC project was highlighted by the White House in its report to Congress as a national priority for the federal 2011 technology budget. The researchers presented their findings at the 6th International Digital Curation Conference, and at the 2010 Joint Conference on Digital Libraries.

As data collections grow bigger, new ways to display and interact with the data are necessary. Currently, TACC is building a transformable multi-touch display to enhance interactivity and the collaborative aspects of archival analysis. The new system will enable multiple users to explore data concurrently while discussing its meaning.

"What constitutes research today at TACC will eventually be integrated into the cyberinfrastructure of the country, at which point it will become commonplace," said Chadduck. "In that way, TACC is providing what I believe is a window on the archives of the future."

Provided by National Science Foundation