# *Challenges of Creating Digital Libraries: Digitizing, Organizing, Storing, and Accessing Content*

*Speakers:*
**Edward Galloway**
University of Pittsburgh
Pittsburgh, Pennsylvania

**Ron Larsen**
University of Pittsburgh
Pittsburgh, Pennsylvania

**Gloriana St Clair**
Carnegie Mellon University
Pittsburgh, Pennsylvania

*Reporter:*
**MIAO Jingang**
Texas A&M University
College Station, Texas

Digital libraries are being used more and more. The three panelists discussed the creation, maintenance, and use of digital libraries.

Edward Galloway reported on the creation of a digital-library infrastructure at the University of Pittsburgh. The University of Pittsburgh Library System (ULS, *www.library.pitt.edu*) began creating digital content in 1998, and its D-Scribe system now hosts 70 digital collections. The sources of content include ULS archives and special collections, faculty and departments, and local cultural-heritage organizations. Among its collections are Pittsburgh history, 19th-century schoolbooks, Chinese monographs, Audubon's *Birds of America*, and George Washington's manuscripts. Galloway showed photographs of items from the collections. The collections are extensive, but the staff is small: two librarians, three support staff, and one to three students.

From 1998 to 2003, Galloway said, the staff of the digital library outsourced most scanning; in 2004, they started scanning documents and photographs themselves. Since 2006, they have scanned almost everything in house with advanced equipment, including DigiBook SupraScan scanners. Challenges have included selecting content, creating descriptive information, tracking physical items, determining specifications, handling requirements, maintaining quality control, and addressing workflow issues.

Ron Larsen, dean of the University of Pittsburgh School of Information Sciences, discussed emerging directions in scholarly publication enabled by advances in computing and communication technology, which he referred to as cyberscholarship. His remarks were based on a 2007 workshop on digital repositories (*www.sis.pitt.edu/~repwkshop*) that was sponsored by the US National Science Foundation (NSF) and the British Joint Information Systems Committee. There have long been theoretical investigations and empirical explorations, but starting 3 or 4 decades ago, high-performance computers have boosted data-driven discoveries: a computer analyzes millions of documents at a time and discovers patterns otherwise undetectable. Some research is communication enabled. For example, there is only one Hubble telescope, and competition to get a few hours with the telescope is fierce. But now, data from the telescope are available online and open to all, and the data that an astronomer needs might already be available.

Most disciplines accept digitalization as the norm, Larsen said, but rather than being digitalized, some primary research data get discarded or are rarely publicly accessible even if saved or published. Therefore, guidelines, norms, and incentives are needed for publishing data and making them publicly accessible. The growth rate of data generation is increasing every year, and the global production of data exceeded production of storage in 2007. The data-management strategy of storing everything and then creating search engines does not work any more, and data curation is necessary. "By innovating value-added solutions, experimenting with new business models, and reaching beyond traditional disciplines, we hope to find a stable equilibrium," said Larsen.

Gloriana St Clair, dean of Carnegie Mellon University Libraries, reviewed developments in open access. She estimated that about 10 million volumes are openly available, 7 million of which are available through Google. The Million Book Project (subset of the Universal Library, *www.ulib.org)* had contributed another 1.5 million as of 2007. The project, which was launched in 2000, was acknowledged by Google Book Search as one of its inspirations. The project is funded by NSF, China, India, and the Internet Archive (a nonprofit organization that preserves Web sites by taking regular "snapshots").

St Clair mentioned that Brewster Kahle, director of the Internet Archive, is concerned about Google's monopoly on some "orphan books" (books that are out of print but still under copyright). She praised the efforts of the National Institutes of Health and Harvard University toward open access. Open access is in the interest of authors, she said, because works available online are more likely to be cited.