Article and Data Repositories

Speakers:
Chuck Koscher
CrossRef
Lynnfield, Massachusetts

Paul Pedersen
Mark Logic Corporation
San Mateo, California

Reagan Moore
National Virtual Observatory
San Diego Supercomputer Center
University of California, San Diego
La Jolla, California

Sean O'Doherty
The Berkeley Electronic Press
Berkeley, California

Reporter:
Rita M Washko
Arizona State University
Tempe, Arizona

Chuck Koscher, director of technology at CrossRef, opened by defining the three common goals of article and data repositories: the preservation of scholarly output of an institution; the dissemination of the information, which advances the image and brand of an institution; and free open access to scholarly content. Depending on one's allegiances, however, the relative importance of the three goals varies. For example, dissemination of information may be most important to those concerned with image, whereas affiliates of the Scholarly Publishing and Academic Resources Coalition (SPARC) would be more interested in free open access.

CrossRef recently convened a subcommittee to explore the institutional-repository phenomenon, said Koscher. One purpose is to define how CrossRef can assist the repositories with their interactions with publishers. Koscher acknowledged that conflict exists between open access to the repositories and the peer-reviewed scholarly system.

CrossRef's subcommittee has found that institutional-repository efforts in the United States lag behind those in Europe, where governments support such initiatives. He noted that institutional repositories in the United States are still in an "embryonic state".

Paul Pedersen, cofounder of Mark Logic Corporation, gave a brief overview of his company before delving into the paradigm of guided navigation. Founded in 2001, the company offers an extensible markup language (XML) search and query system that allows users to seamlessly access and integrate content from various sources, which can then be transformed and republished in the desired format.

Pedersen distinguished the "more primitive process" of conducting a search—in which data are simply concentrated—from his company's system that offers guided navigation, in which any tag can be used as a potential point of access, allowing "fine-grained access to document structure". The process uses predefined categories to fine-tune and classify query results.

Because today's scientists are forming collaborations and need a way for all parties to access the same data, massive archives are required, Pedersen said. As this need for data access increases, so does the need for data storage and manipulation.

"In the future, we'll be publishing scientific data, as well as scientific articles. We're interested in building these links" between publications and the underlying scientific data, said Reagan Moore. The National Virtual Observatory (NVO), funded by the US National Science Foundation, is a collaboration to build a framework for sharing astrophysics data. Thus far, NVO has developed standard terms for identifying astronomy data and standard services that can be used to access a wide variety of existing catalogs and surveys on the Web. Some of the services are a simple image-access protocol, a cone search (a search that locates all objects near a specified location) for catalog records, and a query function for aggregating records from multiple catalogs. NVO is working on additional standard access services, efforts that will improve the efficiency and effectiveness of data-sharing and provide an unparalleled opportunity for discovery, Moore said.

Sean O'Doherty noted that his company wanted to develop technology that would increase scientists' access to information and thus moved into publishing its own journals and developing and marketing software for editorial and repository management to streamline the publishing process. The company publishes 25 peer-reviewed electronic journals, and its technology is used as the backbone of the University of California's eScholarship Repository.

O'Doherty concluded the session by stating that institutional repositories can be seen as giving a new life to materials with a small or nonexistent "preinstitutional repository audience" (conference proceedings, archival data, datasets, sound and video files, interactive presentations, and the like). Those repositories are beneficial because they both lower costs and increase readership, he said. A recent trend is for universities to include postprints, copies of articles by their faculty, in their repositories. The recent National Institutes of Health (NIH) policy strongly recommending that NIH-funded research be made publicly available from noncommercial servers may accelerate the inclusion of postprints in the repositories.