◆ *Archiving Your Legacy: Putting Old Issues Online*

Speakers:
Barbara Gordon
American Society for Biochemistry and Molecular Biology
Bethesda, Maryland

Bernard Stukenborg
Cadmus Professional Communications
Baltimore, Maryland

Bruce Rosenblum
Inera Inc
Newton, Massachusetts

Reporter:
Ted Freeman
Allen Press Inc
Lawrence, Kansas

To the younger generation, Barbara Gordon noted, science not published on the Web does not exist. In moving legacy journal articles online, we bring them back to life and connect them to current research, providing a great service to the scientific community and striking a blow for preservation.

Should we recover all legacy data?

Ultimately yes, said Gordon, who has been involved in creating an online archive for the *Journal of Biological Chemistry (JBC)*. Full-text HTML (based on SGML) and PDF articles from 1995 to the present were already available with free access to all volumes at least a year old. *JBC* now archives volumes from 1980 forward online. The response from *JBC* users has been overwhelmingly positive. *JBC* will eventually include everything since the journal's inception in 1905.

What format should we use?

JBC has chosen a grade of PDF (Adobe's Portable Document Format) called "image + text". That format is widely used because of its affordability and accessibility via the free Acrobat Reader. It is a black-and-white 300-DPI image of the original printed page supported by an underlying uncorrected ASCII text file recovered by OCR (optical character recognition) scanning with around 99% accuracy. The "dirty" ASCII, which is not displayed, provides acceptable searchability. Pages with continuous-tone images receive grayscale scans, and pages with color images are scanned as color.

Step 1 is locating a vendor to do the scanning. Step 2 is locating clean paper copies of the journal and verifying that the set of volumes and issues is complete. Gordon and Bernard Stukenborg warned that we should expect problems with some pages, including poor print quality and occasional dog-eared, stained, or missing pages. We should be prepared to have the spines removed for scanning (they can be rebound if necessary). *JBC* used expendable copies to save money. According to Stukenborg, it is important to have the scanning vendor and Web-hosting provider communicate during the process. The cycle time for the typical legacy conversion project is around 6 months.

Central to creating electronic archives are questions about scope, access, format standards, and the long-term responsibility of the archivist.

How much does it cost?

According to Gordon, *JBC* has spent about $1/page for PDF creation thus far; this does not include the cost of hosting or maintaining the PDFs on the Web site or the cost of creating metadata (SGML titles or abstracts) for database indexing. Stukenborg pointed out that the cost of preserving pages digitally with PDF "image + text" was minuscule compared with the cost of creating a printed page, which two audience members stated was $300 to $400.

How do you pay for it?

Gordon said no cost was charged to *JBC* online journal subscribers, because access to the legacy material is free. Stukenborg suggested that posting legacy articles online increases the original return on investment by extending access to the data. Pay-per-view and separate subscriptions to the legacy material, as some publishers offer, are options for increasing the return even further.

Bruce Rosenblum took a bird's-eye view of the journal legacy question, surveying the evolution of publishing technology from Gutenberg to the Web and the transformations in information flow that online publishing has brought about.

Central to creating electronic archives are questions about scope, access, format standards, and the long-term responsibility of the archivist. Whoever takes on the permanent archivist's role, publisher and archivist will need to agree on access, what rules will apply, and what will trigger them. Having been directly involved in a Mellon-funded study of publisher DTDs (document type definitions) to determine the feasibility of a universal archival XML DTD, Rosenblum stressed the importance of data standards and expressed concern about the proprietary nature of PDF.

What publishers should do now to prepare for archiving, according to Rosenblum, is carefully consider the formats they use to preserve their data, use DOIs (digital object identifiers), and pay closer attention to the quality and accuracy of the data.