A digital preservation primer for scientists

Martin Fenner, Gobbledygook

February 1, 2010

This weeks's blog post is a guest post on the Biomedicine on Display blog – I was kindly invited by Thomas Soderqvist from the Medical Museum of the University of Copenhagen.

> The first email was sent in 1964, but that first email has been lost forever. - Lucy Nowell

As we have moved to digital formats both for primary research data and scientific publications, digital preservation has become critical to secure permanent access to scientific information. Digital preservation turned out to be much more difficult than creating digital content, as preservation requires long-term thinking about many issues including file formats, storage solutions and funding. Digital preservation turned out to be too big for individual libraries, publishers or research disciplines, and large collaborative efforts were started in the last five years.

Alliance for Permanent Access

The Alliance for Permanent Access is a European strategic framework for digital preservation of scientific information. The alliance coordinates the efforts of different funders, research support organizations and major European research laboratories (e.g. CERN or ESA).

Sustainable Digital Data Preservation and Access Network Partners (DataNet)

Sustainable Digital Data Preservation and Access Network Partners is a digital preservation project by the National Science Foundation. The deadline for proposals was May 2009, and $100 million will be awarded over the next five years. Wow.

Portico

Portico is a not-for-profit digital preservation service for scholarly content. Portico was launched in 2005 with initial support by JSTOR, Ithaka, the Library of Congress, and the Andrew W. Mellon Foundation. The Portico archive currently

contains close to 15 million papers and is archiving journal content for many publishers and libraries for a fee. Portico steps in (a so-called trigger event) when a publisher

- stops operations
- ceases to publish a title
- no longer offers back issues
- has a castastrophic failure of the delivery platform

File formats

PDF/A was approved as an ISO standard for long-term archiving of electronic documents in 2005. Before PDF/A, many organizations (including our institution) used the raster graphics format TIFF. The major advantage of the PDF format is the handling text and vector graphics in addition to raster images, allowing full-text search and smaller file sizes. Because the PDFformat is constantly changing, PDF/A was based on a specific PDF version (1.4) with the following specifications:

- self-contained, no external images or fonts
- no sound or movies
- metadata in the XMP format
- no password protection

Most scientific papers are now produced in XML, usually using the NLM DTD. The Archiving and Interchange Tag Set is a flavor of the NLM DTD that is intented for archiving.

Storage solutions

Hard disks, tape and optical media are possible storage solutions. Tape is the ideal solution for long-term storage of research papers, but the digital preservation of research data in many areas (e.g. sequencing, high-energy physics) can't be done with tape because of the exponential growth of these data. Hard disk storage has another problem: the energy requirements of data centers.

We live in a digital world, and this of course includes how we do and communicate science. It is surprising that we have barely started to think about digital preservation.