nature webdebates

PubMed Central decides to decentralize

Edwin Sequeira, Johanna McEntyre and David Lipman

National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
Bethesda, MD 20894, USA
Correspondence to Edwin Sequeira: sequeira@ncbi.nlm.nih.gov

In 1999 the US National Institutes of Health (NIH) suggested that a freely accessible public archive for the scientific literature would greatly benefit the scientific community. To date more than 20 journals have acknowledged this by contributing material to PubMed Central, the electronic archive born from that NIH proposal. In order to encourage wider publisher participation in PubMed Central we will now offer them the option of depositing material for archival purposes without requiring that the full-text of articles be viewable at the PubMed Central site itself. Rather, searches of the archived material at PubMed Central would lead users to the full-text articles displayed at the publisher's site.

Life science publishing, like any other consumer industry, has had to respond to the technological change brought about by the Internet. A large and growing proportion of science journals now publish online versions of their articles, online-only journals are sprouting up and several journals now even take online submissions and orchestrate online formal peer review.

Two years ago, Harold Varmus, then director of the NIH, announced the E-biomed initiative to ensure that there would be a robust electronic archive of life science research articles freely accessible to everyone. After much discussion and the incorporation of the ideas of many people, E-biomed became the free life sciences journal archive PubMed Central. The archive is managed by the National Center for Biotechnology Information, a part of the National Library of Medicine (NLM), USA, which has significant experience in the creation of online archives, exemplified by PubMed (MEDLINE) for biomedical abstracts and GenBank for nucleotide sequences.

Content at the PubMed Central site is submitted by participating journals, which handle article submissions from scientists, peer review and editing in the normal way. The journal articles are supplied to PubMed Central either at the time of publication of the issue, or after a delay of anything from one month to a year or more after publication. PubMed Central archives the articles, and makes them available for readers.

Until now, participating publishers had to allow all articles to be displayed free at the PubMed Central site. The new option removes this restriction. The full text can be searched in PubMed Central or manipulated in other ways, for example, to create links to GenBank resources, but a publisher can now stipulate that full text may only be seen at its own site. The one condition is that these articles must be available free at the publisher's site within no more than one year of publication, and preferably within 6 months. If, at any time, the publisher fails to comply with this condition the NLM will have the right to make the material freely available in PubMed Central a year after publication.

Participating publishers who use this more flexible option will continue to submit information to PubMed Central in the normal way. The text must be in properly formatted SGML or XML files conforming to a publisher's document type definition (DTD) for journal articles. (SGML and XML are standards for marking up the parts of a document. A DTD describes a specific mark-up template.) All necessary high-resolution image files must also be supplied. For journals that choose to make their material viewable in PubMed Central, PDF files and supplementary data files may also be included. Whether submitted articles are viewable in PubMed Central or not, it is important that they meet a PubMed Central standard for completeness and syntactical correctness so that the integrity of the archive is ensured.

Over the past few months PubMed Central has developed a new software architecture for the archive, built around the concept of a common template and a precise specification for data tagging. This new PubMed Central DTD, based on the latest XML standards, creates a more detailed and sustainable archival copy of an article than HTML, which currently serves as the "standard" electronic record for most online journals. Figure 1 illustrates how the archive is created.

The common PubMed Central DTD means that every article in the archive has its parts (authors, affiliations, major article sections, references, etc.) tagged in exactly the same way, regardless of its source SGML or XML format. A measure of the diversity of tagging schemes in use today is that the first 25 journals submitting data to PubMed Central use 10 different DTDs. PubMed Central's normalized tagging, which has no effect on the article's content, greatly simplifies all further use of the archive.

For example, uniform tagging allows searches for reagents mentioned only in the methods section of an article or searches of just the figure and table legends. In fact, it simplifies the provision of any feature that depends on knowing the context of a string of terms in an article. The further integration of the content of journals participating in PubMed Central with other NCBI resources -- such as whole genomes, macromolecular structures and online textbooks -- is also facilitated by a consistent data structure within the archive. The process of creating honed search engines and adding more value to PubMed Central articles through linking to related information will benefit from access to the same tools that have been used to create PubMed, GenBank and other NCBI resources.

As with any new kid on the block, it is true that we may not yet have as broad a set of capabilities as the more established players. In our recent efforts, much of the emphasis has been on the distinctly unglamorous but absolutely essential work of creating a stable and robust archive architecture. We are set to build and to make big strides forward in this area. Furthermore, by virtue of starting later, we have had the greater flexibility that enables us to adopt newer technologies and, therefore, bring completely new capabilities to the playing field.

The National Library of Medicine has been collecting and preserving the medical literature for more than a century. Extending this stewardship to the electronic literature is a natural and responsible role for the NLM to play. But what does preservation have to do with free access? The reality is that the only way to assure the permanence and durability of an electronic archive is to use it continuously and there is no better way to do that than by making it freely available to everyone. So, PubMed Central was built on the twin standards of a permanent archive and free access. The latter is now being stretched to allow users to be directed to a publisher's site for full text. We hope that the new flexibility will encourage many more publishers to contribute to the archive so that it can realise its full potential -- in ways that are still to be discovered.

Rules of the Game

The Basics

- Journal supplies full-text data and accompanying images that meet the PMC standard for completeness and syntactical correctness.
- PMC maps incoming SGML/XML to the PMC DTD format. This mapped data, along with associated images and other files, makes up the PMC archive.
- Full text of all deposited material may be searched in PMC to produce a list of qualifying articles.

If Journal Makes Full Text Available in PMC

- Full text is viewable free in PMC, although a journal may embargo release of its material in PMC for weeks or months after publication.
- The journal banner on every journal page (table of contents, abstract, full text, etc.) in PMC includes a link back to the journal's own site.
- If the journal is not already in PubMed, NLM adds citations to PubMed for the articles available for display in PMC.

If Journal Makes Full Text Available Only at Its Site

- When an article is included in a search result, PMC will provide a link to the full text at the journal site instead of presenting the full text in PMC.
- All material to which PMC links at the journal site must be available free and without access restrictions within no more than one year of publication, and preferably within 6 months after publication. If, at any time, the journal fails to comply with this condition NLM will have the right to make the material freely available in PMC one year after publication.

Dispelling the Myths

The inclusion of pre-print material and unreviewed research reports taints content provided by journals with rigorous peer review standards.

In fact, from the day it went public in February 2000, PubMed Central has explicitly excluded pre-prints and research reports that have not been peer reviewed. These classes of literature were excluded because of the strong objections of publishers to the original E-biomed proposal. This policy will not change.

Journals risk a loss of quality when they allow their content to be displayed in PMC. PMC may not display information correctly and could introduce errors.

The initial setup for a journal in PMC includes a thorough automated and manual review of the journal's content to ensure that the accuracy of the material as presented in PMC is at least as good as that on the journal's own site. A publisher also gets to review the content in PMC before it is released publicly. These reviews and continuing reviews of new content actually enhance the quality of a journal's electronic archive by conducting an independent test of its usability. All submitted articles must meet a PubMed Central standard for completeness and syntactical correctness to ensure the integrity of the archive. If errors are found during the quality control process the journal is asked to provide correct data. Conversion to the PMC DTD format normalizes the tagging of the article's parts (authors, affiliations, major article sections, references, etc.) but has no effect on its content.

Participating in PMC involves some kind of "exclusive use" agreement.

The archiving of electronic content in PMC is no more exclusive than NLM's preservation of the printed literature. In fact, NLM advocates multiple copies of the archive managed by other organizations in order to better ensure the durability and continuing usability of the archive. In addition, NLM stands ready to give a publisher a copy of its content whenever requested.

The cost of sending data to PMC is too high.

Journals that already produce SGML or XML versions of their articles incur, at most, a nominal cost from their production vendors for sending data to PMC. PMC itself charges nothing for archiving the content or for any related services.

click here to see figure and legend