PubMed Central decentralized

Sir — 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 contributed material to PubMed Central, the electronic archive born from that NIH proposal. To encourage wider participation, PubMed Central (http://www.pubmedcentral.nih.gov/) is now offering publishers the option of depositing material for archival purposes without making the full text viewable. Searches for this material at PubMed Central would lead users to the full-text articles 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, growing proportion of science journals now publish online versions; online-only journals are sprouting up; and several journals now take online submissions and use online peer review.

Two years ago, Harold Varmus, then director of the NIH, announced the E-biomed initiative to ensure 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 with diverse interests, E-biomed became PubMed Central, the free life-sciences journal archive. It is managed by the National Center for Biotechnology Information (NCBI), a part of the US National Library of Medicine (NLM), which has significant experience in the creation of online archives, exemplified by PubMed (MEDLINE) for biomedical abstracts and GenBank, for nucleotide sequences.

As things stand, participating journals submit their content to PubMed Central either at the time of publication or after a lag of up to a year or more after publication. PubMed Central archives the articles and makes them universally available to readers. Until now, participating publishers had to allow their full text to be displayed free at the PubMed Central site where it could be searched or manipulated, for example to create links to GenBank. Our new option means that a publisher can now stipulate that full text is seen only at its own site, with the publisher's own restrictions on access. PubMed Central's sole condition is that this content is made available free at the publisher's site within one year of publication (preferably within six months). If the publisher fails to comply, the NLM will have the right to make the material freely viewable in PubMed Central a year after publication.

Publishers choosing the new option will submit their full text to PubMed Central as they do now, in SGML or XML (mark-up languages) files conforming to a document-type definition (DTD — a mark-up template) for journal articles, and as high-resolution image files. Viewable and non-viewable submitted articles must meet a PubMed Central standard to ensure the integrity of the archive (see 'rules of the game' in our contribution "PubMed Central decides to decentralize" at http://www.nature.com/nature/debates/e-access/) .

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

The new standard DTD means that every article in the archive has its parts (authors, affiliations, text, references, and so on) tagged in exactly the same way, regardless of its source format. (The first 25 journals using PubMed Central have 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. It allows, for example, searches for reagents mentioned only in the methods section of an article or searches of just the figure legends. It simplifies the provision of any feature that depends on knowing the context of a string of terms in an article. Further integration of the content of PubMed Central journals with other NCBI resources — such as genomes, macromolecular structures and online textbooks — is also facilitated.

As with any new kid on the block, we may not yet have as broad a set of capabilities as the more established players. Much of our recent emphasis has been on the unglamorous but essential work of creating a stable, robust archive architecture. By starting later, we have greater flexibility to adopt newer technologies and therefore to introduce new capabilities.

Many myths have arisen round PubMed Central. It does not publish preprints and other unrefereed material; participation does not involve any

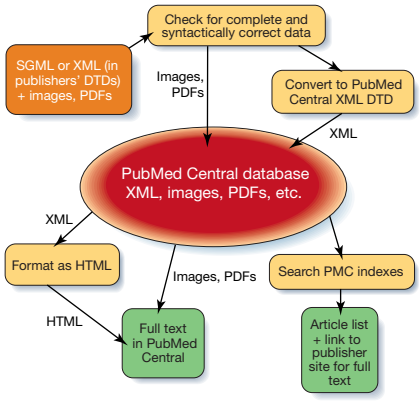

Figure 1 **PubMed Central data flow.**

'exclusive use' agreement; and PubMed Central does not charge for archiving or related services (see 'Dispelling the myths' in "PubMed Central decides to decentralize" at http://www.nature.com/nature/debates/ e-access/ for more details). The NLM has been collecting and preserving the medical literature for more than a century; the extension to stewardship of the electronic literature is a natural step.

But what does preservation have to do with free access? The only way to ensure the permanence of an electronic archive is to use it continuously, and there is no better way to do that than making it freely available to everyone. PubMed Central was built on the twin standards of a permanent archive and free access. We are now stretching the latter principle to direct users to a publisher's site for full text. We hope that many more publishers will thereby be encouraged to contribute to the archive so it can realize its full potential — in ways still to be discovered.

Edwin Sequeira, Johanna McEntyre, David Lipman
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA

● This article is a slightly edited version of a contribution to *Nature*'s current web debate on electronic publishing initiatives in science (see *Nature* **410,** 613; 2001). Readers wishing to participate in the debate by replying to this or any of the other contributions are invited to view (http://www.nature.com/ nature/debates/ e-access/). Contributions can also be submitted to Correspondence via corres@nature.com. In either case, publication will be offered according to the criteria described on page 613 of last week's issue — Correspondence Editor, *Nature*.