Preservation of Knowledge, Part 2: Digital Archives

M. Castillo

AJNR Am J Neuroradiol published online 2 July 2009
http://www.ajnr.org/content/early/2009/07/02/ajnr.A1656.citation

Published July 2, 2009 as 10.3174/ajnr.A1656

PERSPECTIVES

Preservation of Knowledge, Part 2: Digital Archives

In last month's *Perspectives*, I addressed the role of paper and microfilm as media for the preservation of written knowledge, particularly biomedical literature. A significant number of journals are now electronically archived including the *American Journal of Neuroradiology* (*AJNR*). The HighWire Press data base contains 1232 journals and more than 5.6 million articles as of this writing. According to my last count, the Elsevier and Springer journal data bases contain 2320 and 2084 scientific journals, respectively! It is hard to believe all of these journals can be kept in their past and current print forms and archived for posterity. Most current journals are, however, archived in digital form. The main topic of this article is to discuss some of these electronic data bases.

Although one of the major goals of libraries is to avoid duplicated material and save space, one of the primary goals of electronic depositories is to create redundancy and several copies of their contents. The *AJNR* is preserved by a system called LOCKSS (Lots of Copies Keep Stuff Safe).[1] This system is based on technology designed by the Association of Computing Machinery. The LOCKSS system is, like the HighWire Press, based at Stanford University and provides libraries with open-source software to preserve all sorts of materials that have been published on the World Wide Web. Each library owns a LOCKSS "box" that allows perpetual access to their materials. LOCKSS has a self-checking mechanism that continuously audits and repairs the information it houses. This is accomplished by "crawlers" that compare the LOCKSS box contents with an institution's Website and constantly update the box contents.

The *AJNR* allows these crawlers on our Website where information is collected and used to update our LOCKSS files. However, unlike libraries, we do not administer this mechanism; HighWire does. Decentralization of the system assures its independence from central failures by creating multiple archives in different locations and constantly comparing these with replicas. Today, approximately 400 publishers and hundreds of libraries worldwide use LOCKSS. The next iteration of this system is CLOCKSS (the C stands for "controlled"). Through it, the main research libraries in the world and major publishers are creating a "dark" archive of all of their contents to further assure the preservation of digital data. LOCKSS can also be used to keep older literature that has been digitized. Because software and hardware are in a state of constant evolution, data may "rot," and all preservation systems need to be constantly updated.

In June 2004, the Wellcome Trust, Joint Information Systems Committee, and the US National Library of Medicine announced a joint effort to digitize back files of what they considered to be important medical journals.[2] This effort is known as the Medical Journals Backfiles Digitisation Project and is one of a total of 6 such projects in the world. All of the digitized files, some dating back 125 years, became open access. This data base does not contain any US-based imaging-related journals, but a search revealed that the *Korean Journal of Radiology* is included.[3] In addition, some radiation oncology journals are found there.

Digital archiving initiatives are vast. A complete list may be found at the National Digital Information Infrastructure and Preservation Program of the Library of Congress.[4] Other countries have also joined in this effort. The National Library of the Netherlands (KB) has created an "e-depot" system available to all publishers whose main goal is to maintain the integrity of digitally stored objects. This depository is neither "dark" nor "light," but each user has access only as established by a previous individual agreement. All publications contained in KB's e-depot that come from BioMed Central continue to be open access. The storage capacity of this site is expected to reach 1.5 petabytes soon. This endeavor is closely associated with other European ones such as Preservation and Long-Term Access through Networked Services (www.planets-project.eu), the purpose of which is to build practical services and tools to ensure access to digital culture and scientific assets. Another interesting and more far-reaching project is Cultural, Artistic and Scientific Knowledge for Preservation, Access and Retrieval (www.casparpreserves.eu), which preserves all kinds of digital data in a technology-neutral, domain-independent centralized system to assure its longevity. In the United States, another huge project is Portico (www.portico.org), which is sponsored by the Library of Congress and other nonprofit organizations. The Portico Website lists some interesting facts stressing the importance of digital archiving as follows:

- A total of 13% of articles cited by the *New England Journal of Medicine, Journal of the American Medical Association,* and *Science* are irretrievable from the original hyperlink only 27 months after publication; hence archival of all materials is important as their on-line longevity may be very short.[5]
- In 2002, a total of 70% of faculty in a survey were using electronic journals for research, and 1 year later, nearly 80% considered these as "invaluable research tools."[6]
- Like other digital archiving initiatives, Portico follows the standards set by the National Library of Medicine Journal Archiving and Interchange Suite.[7] These standards set the way documents are saved and transferred. Portico houses information from 487 libraries and 58 publishers (including Elsevier and Springer), including more than 11 million articles. Many other digital-archiving initiatives (JSTOR [Journal STORage at the Mellon Foundation], Ithaka, Aluka, *Journal of the American Medical Association* and *Archives* Journals Backfiles, etc) are available, but I cannot mention all of them in this *Perspectives* because of space limitations.

Digital archiving extends beyond science to all of our daily activities. From the movie industry to our own home videos and photographs, from the nation's digital memory to that of persons, digital archiving is of significant importance. E-futurists assure us that, in the future, all written material will be found in an electronic form and probably will be carried on portable devices. Other imaging-related journals (eg, *Radiol-*

EDITORIALS

ogy and the *American Journal of Roentgenology*) use HighWire and entrust their archives to LOCKSS. Our readers can be assured that the digital contents of *AJNR* are being adequately preserved for future generations.

References

1. **LOCKSS.** Available at: http://www.lockss.org. Accessed March 30, 2009
2. Mehnert R, Cravedi K. **International Agreement to Expand PubMed Central. US Library of Medicine, National Institutes of Health.** Available at: http://www.nlm.nih.gov/news/press_releases/intlpubmed04.html. Accessed March 30, 2009
3. **PubMed Central Journals.** Available at: http://www.pubmedcentral.nih.gov/fprender.fcgi. Accessed March 30, 2009
4. **Digital preservation.** Available at: http://www.digitalpreservation.gov. Accessed on March 30, 2009
5. Dellavalle RP, Hester RJ, Heilig LF, et al. **Information science: going, going, gone: lost internet references.** *Science* 2003;302:787–88
6. Portico. **Why Archive?** Available at: http://www.portico.org/about/why.html. Accessed on April 6, 2009
7. **National Library of Medicine Journal Archiving and Interchange Suite.** Available at: http://dtd.nlm.nih.gov. Accessed March 30, 2009

M. Castillo
Editor-in-Chief

DOI 10.3174/ajnr.A1656