SECTION 108 STUDY GROUP
WRITTEN COMMENTS ON TOPIC 4

Name: William Y. Arms
Professor of Computer Science, Cornell University

Organization: Faculty of Computing and Information Science, Cornell University

Contact: Cornell Information Science
301 College Avenue
Ithaca NY 14853

New Website Preservation Exception

The Web is one of the most important cultural artifacts of our time. For social scientists especially it is both a subject of study for itself and a source of evidence about contemporary social events. Yet, even in its brief history, enormous amounts of Web content have been lost for ever. Partly this is the result of the current legal framework, which inhibits archiving and preservation.

These recommendations are based on practical experience in several large projects that have included collection and preservation of Web materials. The specific projects are as follows:

- The Cornell Web Library and Laboratory. This is a large-scale project funded by the NSF to organize the historical collections of the Internet Archive for scholarly research. These collections contain crawls of the open-access Web taken every other month, since 1996. [For an overview see: http://www.dlib.org/dlib/february06/arms/02arms.html.]

- The Library of Congress's Minerva Web archiving program (consultant 2000-2001). [See: http://lcweb2.loc.gov/cocoon/minerva/html/minerva-home.html.]

- The NSF's National Science Digital Library (NSDL). This large-scale program is building a digital library for all of science education. This program began with collection building and the organization of existing materials, but preservation has become a vital part of the long-term success. [The NSDL is at: http://nsdl.org/.]

Recommendation 1. The rules that govern Web preservation must be technically flexible and not make assumptions about today's technology nor how it is likely to change in future.

Currently, preservation strategies for the Web fall into two categories: selective and automated. With selective preservation, decisions on the sites to be collected, frequency of collection, etc. are made explicitly by librarians. Minerva is an example of this approach. With automated preservation, there is no selection of individual sites and minimal metadata. The Internet Archive has been the leader in this work.

Both approaches are important, but the Web preservation exception should not assume that these are the only possible approaches. The techniques for archiving and preservation will continue to change. For example, there is promising research into methods of expert-guided selection and collection, which combine human expertise with automated techniques based on machine learning and natural language processing.

Recommendation 2. A method should be developed for Web site owners to indicate their wishes with respect to the collection of the materials for preservation and access after preservation.

The general principle behind these recommendations is that unless a Web site explicitly states otherwise, any library or archive can collect open-access content, store it for preservation, and make it available to scholars.

A mechanism should be developed that would allow Web sites to indicate exceptions to the general principles. This mechanism could be an extension of the robots.txt standard, or a separate parallel system, perhaps a preserve.txt file. The label might list which libraries and archives can or cannot collect the material, and access restrictions after collection. A valid email address for the copyright owner would be required.

Recommendation 3. The Web preservation exception should explicitly permit libraries and archives to delegate all operational aspects of preservation to other organizations.

Eligibility is the subject of Discussion Topic 1 and not discussed here.

Large-scale collection of digital materials poses policy questions beyond copyright. For example, Web collections raise complex questions of curation and privacy.

The Internet Archive, the Cornell Web Library, and the NSDL are examples of new types of organizations that have an important role to play in preserving Web information. One reason for the emergence of new types of organization is that many of the methods for collecting and preserving Web data require specialized expertise, beyond that which is found in libraries and archives.

At the very least, the changes to Section 108 should permit libraries and archives to delegate this technical work to other organizations. These organizations might be not-

for-profit, such as the Internet Archive, which currently acts as a contractor to several national libraries or commercial organizations. In this context, the term "delegate" rather than "outsource" is important.

Recommendation 4: The Web preservation exception should apply, without special authorization, to all materials on the Internet that are accessible by a person using a Web browser or by a computer using standard Web protocols.

By placing material on the Web with open access, the copyright owner indicates an expectation that the general public will read and make use of the materials. Unless materials are labeled otherwise, it can be presumed that the copyright owner is willing to have limited use made without specific authorization, including copies made for preservation.

This definition includes materials that are labeled with machine-readable guides to use (e.g., robots.txt files), or human readable copyright licenses.

Recommendation 5. Provision should be made to collect and preserve all Web materials, not withstanding exclusions such as robots exclusion and requests from the copyright owner.

Preservation of Web content by third parties, such as the Internet Archive or NSDL, is limited by robots.txt exclusions and other requests from copyright owners not to collect their pages. Respecting these restrictions means that some of the most valuable materials are being lost. Newspapers and state government documents are prime examples.

On the assumption that copyright is a balance between the interests of the copyright holder and the public good, there should be a straightforward mechanism to allow all material that is made available on the Web to be collected and available for scholarship. At the very least some variant of mandatory deposit is needed to allow the Library of Congress or its agents to collect all materials on the Web, for long-term preservation and access. For redundancy, there should be several independent agents.

These procedures must not be burdensome to either party and must be capable of being automated. In particular, there should not be a requirement to obtain explicit permission from each Web site owner.

Recommendation 6. Libraries and archives of Web materials should have acceptable use policies that all scholars must accept before using the archives. The policies for access to preserved Web materials should be expressed in terms of the category of use and respect for the interests of the copyright owners.

The topic of access to preserved digital information is extremely complex and is the subject of Discussion Topic 2. The policies for Web archiving can be simplified because all the material has been made available with open access. Most of the material that is placed on the Web with open access has no commercial value six months after it is

posted. However, some Web materials are commercially valuable and it is important to have procedures to protect them.

Unless the copyright owner has explicitly labeled the materials as commercially valuable over the long-term, all libraries and archives should be permitted to develop policies that allow non-commercial use by scholars six months after collection. Use that is compatible with the principles of fair use should always be permitted. In particular, the Library of Congress should have a carefully monitored set of procedures that allow use by scholars of its entire collection.

Web data is intrinsically digital information designed to be delivered over networks. Policies about access should anticipate that most use of Web archives will be by remote users who connect over networks. Scholarly research is changing, with an ever-increasing amount of automated processing. The policies should expect that researchers will download copies of Web materials to their own computers for analysis.

Footnote: technical considerations

There are no technical barriers to implementing these recommendations.

Modern Web crawlers have *politeness* algorithms that can ensure that the demand that a single crawler places on a Web site is not excessive. However, it may be necessary to restrict the number of crawlers that access a site for preservation. There should also be a naming convention so that all crawlers that claim the exception clearly identify themselves.

The question of how to identify and collect all the software that is needed to preserve the full user experience of digital information is probably not solvable in general. However, this problem is less severe with Web information than with other digital information, since most Web pages can be rendered with a standard browser and a limited set of plug-in modules.

William Y. Arms
April 19, 2006