Comments on
Exceptions and Limitations Applicable to
Libraries and Archives Under Section 108 of The Copyright Act
Notice of Inquiry, Copyright Office, Library of Congress
April 10, 2006

Patricia Cruse
Director, University of California Libraries' Digital Preservation Program
on behalf of
California Digital Library
University of California Libraries

The University of California system is comprised of ten research institutions. The collections of the University include a coordinated 10-campus library system with holdings of over 34 million bound volumes, nearly 200 million manuscript items, 2 million maps, a host of resources in other formats, and over one thousand museums, galleries, arboreta, herbaria, and archives that hold more than150 million objects, including written works, art objects, cultural artifacts, and scientific specimens. The UC libraries play a pivotal role in supporting the University's world-class research, teaching, and service programs through the services they provide on the ten campuses, through systemwide collaborative programs and services and a variety of collaborations with other collecting organizations throughout the UC system. In partnership with the UC libraries, the California Digital Library (CDL) provides much of the Universitywide infrastructure that coordinates services across the system, leverages campus resources and investments, and provides systemwide access to the University's rich information resources.

I am the Director of the UC Libraries Digital Preservation Program, which was established to ensure long-term access to the digital information that supports and results from research, teaching, and learning at UC. The program includes ejournals, web-based content, digitally reformatted materials from UC libraries and museums, and online teaching and learning materials. The California Digital Library is also the recipient of an NDIIPP grant, The Web at Risk: A Distributed Approach To Preserving Our Nation's Political & Cultural Heritage. This grant is enabling us to build a Web Archiving Service that will allow the UC libraries to continue their historic role of collection building in a web-based environment.

The ability to continue to build and preserve collections in a digital environment is of critical importance to libraries and archives, which play a unique and vital role in providing long-term access to and use of cultural, historical, scientific, and financial material. College and research libraries in particular are integral to the mission of higher education by supporting teaching, learning, research, and the creation and dissemination of knowledge. New technologies provide opportunities for libraries to

more broadly and effectively fulfill their primary role of providing access to collections over time.

The explosion of digital information is well documented. The impact of digital information is now firmly implanted in the academic community, and is a core component of academic research, learning, and teaching. Libraries have embraced the use of digital content as part of the way they do business. Library users have also embraced digital information and rely on the availability of information delivered to their desktops 24/7. College and university libraries utilize digital technologies to support the teaching and research needs of scholars and instructors, and to provide digital access to material that previously required travel to the particular library that owned the item. Copyright law must allow libraries to continue their historic mission of collecting, managing, preserving, and providing access to digital information. This statement will address the following areas of inquiry:

- **Topic 1: Eligibility for Section 108 Exceptions**
- **Topic 3: New Preservation-Only Exception**
- **Topic 4: New Website Preservation Exception**

Topic 1: Eligibility for Section 108 Exceptions

It is evident from our experience that limiting Section 108 to “libraries and archives” no longer encompasses the range of collecting organizations that (a) hold valuable and often unique research materials that are protected by copyright, and (b) could make a legitimate and valuable contribution to research if afforded the Section 108 exceptions. In providing service to researchers and students, the UC libraries rely on arrangements that include virtual libraries, such as the CDL, shared print and digital collections and repositories, and partnerships with third parties, such as the Open Content Alliance. The benefits of these collaborations can be unreasonably stymied if the limited exceptions afforded by section 108 apply to only some of the partners.

We believe, therefore, that there is an important distinction between the nature of the 108-sanctioned activity and the nature of the organization engaged in that activity. In this regard they key is subsection 108(a) list of criteria that must be met in order to take advantage of the exceptions. The criteria provide adequate protection for rights holders while avoiding the potential morass of definitional problems in creating an exhaustive list of organizations that might, at any particular time under any particular technology and public service regime, legitimately use the Section 108 exceptions. In short, the key is in the activity - the effective management and control of the copies under current criteria - not in the nature of the organization making the copies.

Definition of “Libraries” and “Archives”

The Section 108 Study Group appropriately recognizes that the lack of definition creates possible ambiguities. However, we would prefer language that clarifies the importance

and prescribed activity permitted under section 108, thereby concentrating on the appropriate use of the section 108 exceptions rather than the nature of the organization invoking them. We would not want to see the addition of language that would narrow the definition of libraries and impede them in fulfilling their historic mission.

Eligible Institutions
Eligible institutions should not be limited to nonprofit and government entities for some or all of the provisions of section 108. As stated above, any narrowing of the language in section 108 would unnecessarily restrict legitimate Section 108 activities due to inevitable oversights and incomplete organizational definitions. Due to the rich expansion of organization partnerships being built to support their historic missions, such narrowing by "nature of organization" would also impede libraries and archives in fulfilling their historic missions.

Non-physical or "Virtual" Libraries and Archives
The University of California system comprises 10 research institutions, a coordinated 10-campus library system, millions of holdings, and thousands of collections. The UC libraries play a pivotal role in supporting the University's world-class research, teaching, and service programs. In partnership with the UC libraries, the California Digital Library (CDL) provides much of the Universitywide infrastructure that coordinates services across the system, leverages campus resources and investments, and provides systemwide access to the University's rich information resources.

The CDL is a virtual library of the University of California, providing digital content for the teaching, research, and learning needs of the 10 UC campuses. The CDL works with the UC libraries to play a pivotal role in supporting the University's world-class research, teaching, and service programs through the collections (digital and analog) and services it provides on all campuses.

Including non-physical or "virtual" libraries or archives within the ambit of section 108 would benefit the CDL and UC libraries in several ways:

- **Continuing our mission and sharing unique resources.** Virtual libraries allow us to continue our historic mission in a digital environment: acquiring content, managing content, providing access to the content, and preserving the content for future teaching, learning, and research. Virtual collections such as the Online Archive of California provide the ability to share valuable and unique resources beyond our own library walls.
- **Reducing costs:** CDL's shared digital collections comprise more than 12,000 journal titles and 250 databases, as well as other materials. This allows campuses access to publications online that they would not have purchased in print due to price constraints. This is a significant organizational innovation in collaborative collection development, allowing the campus libraries to act as a single entity. In a time of rising costs and shrinking budgets, virtual collections drive down the costs associated with the acquisition, management, and preservation of digital content. (For example, an analysis of the top 11 digital

journal publishers showed that consortial purchasing by the CDL results in a 58 percent discount from the average print subscription price, and a gain of more than 13,000 additional subscriptions systemwide. Had these additional subscriptions been purchased by the campuses in print format at list prices, their cost would have exceeded $25 million.)

- **Collaborative environment for resource preservation.** The UC Libraries Digital Preservation Repository lowers barriers, allowing libraries to continue their historic mission. The CDL builds the technology, and campuses use that technology to collect, manage, and preserve content. This same environment can be extended to a broader community, enabling under-funded institutions to make their content available. (For example, the CDL and the UC libraries work with many public libraries and historical societies around the state to make their content available online.)
- **Responding to the changing nature of education.** Distance learning and online education and research have found a place alongside traditional education and research modalities. Virtual collections meet this growing user demand for digital access.

Expand the Scope of Section 108 to Include Museums

The scope of section 108 should be expanded to include museums. Materials generated by museums are increasingly important to the research, learning, and teaching on the UC campuses. The CDL and UC libraries are engaging more and more in partnerships with UC museums to make their valuable materials available. The Online Archive of California <http://www.oac.cdlib.org/> hosts an abundance of materials from the museum community.

Other Types of Institutions

Other types of institutions should also be considered for inclusion in section 108. The challenges and scale presented by digital information mean that libraries must be able to respond as necessary to continue their historic mission. Often this means that the solution includes a third party. Third parties, or third party contributions to library activities, must meet the criteria for appropriate section 108 use. Third parties do not undermine our intention and defensible record of providing every protection for the involved materials.

For example, the UC libraries collaborate with the Internet Archive (a non-profit organization that preserves web content by taking “snapshots” of web sites) to digitize public domain materials from the UC collections for the Open Content Alliance. This collaboration allows UC facilities to host scanning workstations owned and operated by Internet Archive. Although these scanning operations are currently limited to out-of-copyright materials, it might make sense for UC to use the same high-volume facilities and operations to scan in-copyright material (including the creation of copies permitted under section 108, as long as the creation and use of those copies was compliant with applicable law). In our view, the important distinction between in-copyright and out-of-

copyright material in this context is in the effective management and control of the copies, not in the nature of the organization making the copies.

The California Digital Library, as a virtual library, would use the same strict criteria as the UC libraries to assure security for materials, whether handled on the premises or elsewhere.

Topic 3: New Preservation-Only Exception

A new preservation-only exception has important implications for cultural preservation. Making digital copies—most of them ephemeral and inaccessible to outside users—is a natural side-effect of the activities that support preservation. Limiting the number of such temporary copies is infeasible to specify, let alone implement, but techniques such as cache flushing and securing of local networks provide adequate protection to rights holders. We also advocate creating "preservation copies" for to be used only in the event of failure of the original. The preservation copies would not be made accessible otherwise.

Eligible Institutions

We firmly believe in the need for an exception that would permit libraries to preserve digital assets that are key to fulfilling our historic mission. Libraries, archives, and museums seek to preserve the items in their collections in order to protect their investments (frequently, investments of the public's funds) in those resources and to ensure their ongoing availability for research and teaching, consistent with their organizational missions. Because all such institutions share a similar imperative, there should not be statutory limitation that grants this exception only to a select group of institutions. Instead, limits should be implemented in the form of standards and best practices that are developed and agreed upon by the community.

Up-Front Preservation

The inherent nature, volatility, and fragility of digital materials makes a compelling case for up-front preemptive preservation for all digital content. Preemptive digital preservation is crucial element in the lifecycle of digital information. In the world of physical objects, preemptive preservation is not necessary because print conservation techniques can mitigate damage. A damaged book can sit on a shelf for years, and still be both usable and reparable. In a digital world, however, there is no such thing as "partially damaged." Once damaged, a digital file is unusable and most likely irreparable. In addition, published printed information often exists in multiple copies, affording the possibility that some other copies may remain available if a library's local copy becomes unusable. There is no such guarantee for digital materials.

It is likely that a new exception will be required to address the issue of up-front preservation of digital materials. The provisions of subsection 108(c) are based on

physical materials. Attempting to revise this section to accommodate, and spell out conditions for preemptive archival copying of digital materials risks a result that is both inadequate for the digital environment and compromises the proven effectiveness of this provision in the print world.

We take seriously the job of building systems that protect our investments (that is, our digital assets), and we extend that same protection to rights holders. Technical protections are supported by policies, best practices, and guidelines designed to assure that the technical solutions are sound and adhered to.

Technical solutions are a large part of guarantying the rights of content producers, and "trust" is an additional component of digital repositories. Specifically, trust that digital repositories are capable of reliably storing, migrating, and providing access to digital collections. The digital preservation community is looking to the work of the joint task force between RLG and the National Archives and Records Administration (NARA) on the certification of preservation repositories. These experts were asked to define certification requirements, to delineate a process for certification, and to identify a certifying body (or bodies) to implement the process. They have produced a draft report, Audit Checklist for Certifying Digital Repositories, which will assist in determining whether a digital repository can be certified as a trusted location for digital collections. These efforts need to mature in the community.

At-Risk Designation

The preservation only exception should not apply only to a defined subset of copyrighted works, such as those that are "at risk." An "at-risk" designation would be far too ambiguous and be open to too many interpretations by individual institutions. Institutions should be allowed to preserve materials that support the mission of their institutions regardless of the items' perceived "at-risk" attributes. Further, there should be no statutory limitations on "at-risk" materials. An "at-risk" designation would be open to too many different interpretations by individual institutions. Rather, institutions must have the capacity to preserve materials that support the mission of their institutions.

"Dark" Archives

The Study Group asked whether copies made under a preservation-only exception should be maintained out of circulation in proven restricted (or "dark") archives. Currently, a great deal of investigation and debate is taking place in the digital preservation community surrounding the issue of "dark" archives. Now, however, is not the time for legislation. Rather, standards and best practices should emerge from the community.

In addition, current debate is recognizing that the long-term health of digital content is difficult to guarantee and requires inputs from a variety of stakeholders. Automated solutions such as integrity checking (check-sums) can report on the technical health of object, but human use and access can help guarantee the digital content's health and long-term viability.

Topic 4: New Website Preservation Exception

The website preservation exception has broad implications for the University of California's ability to continue its historic role in collection building.

As content producers move their publication processes online and away from traditional print media, some of our ability to acquire and provide long-term curation for materials essential for research (such as government documents) has been compromised. Web-based information is an increasingly critical part of our nation's heritage. This information is fundamentally at risk due to the lack of systematic preservation and the extremely expensive and uncertain process of rights-clearing. We believe that it is imperative for research, public, and special collection libraries to be able to provide enduring access to these essential documents in ways that are cost-effective and that protect the interests of rights holders. Our recent experiences in website preservation with NDIIPP funding from the Library of Congress confirm our view that rights clearing is one of the most significant challenges for web preservation.

Online Capture and Preservation
A special exception should be created to permit the online capture and preservation by libraries and archives of certain website or other online content. Information about our nation’s social, scientific, and political life plays a fundamental role in our society, and public access to it is a basic foundation of democracy. This information is enormously diverse, produced by a wide variety of public, private, and government institutions, and serves multiple audiences, including research institutions, business enterprises, and private citizens.

Memory organizations, government agencies, legal entities, and society as a whole rely on this information for a variety of essential civic, economic, and political functions, and they need access to all types of information, regardless of format. Memory organizations have historically served to preserve these printed materials. Selected state, public, and academic libraries already preserve printed government materials—nearly 1,300 libraries nationwide participate in the Federal Depository Library Program. These libraries maintain collections that extend back to the origins of American governments, and are uniquely positioned to ensure continuity in this historic record as government materials transition from print to digital formats.

Today, however, documentation of our social and political history, and publication of materials that support technical and scientific research is not limited to traditional print publication and dissemination mechanisms. A great deal of material is solely digital. These digitally published materials are inherently volatile, uncontrolled, and at great risk of being lost. The most volatile and at-risk information is that made available exclusively via the World Wide Web.

Memory organizations, particularly libraries, have a central role to play in preserving this web-based content. Memory organizations have long-standing experience in managing and supporting the use of comprehensive collections, and in organizing those

collections to meet their users' diverse needs. In fact, memory organizations are among those few agencies now asserting themselves as the guardians of governments' web-based outputs.

It is crucially important that libraries continue their collection-building efforts by capturing, managing, and preserving web-published materials. With funding from the NDIIPP program, the California Digital Library and its partners are building a web-archiving service that will enable libraries to build collections of web-based materials. We recognize that there are new overlapping classes of web content: 1) web content that is event based and disappears quickly (e.g., bulletins on Hurricane Katrina); and 2) web content that is more stable and is routinely collected (e.g., agency web pages.) Although these are different and demand different strategies, both require libraries to freely collect materials that meet the mission of their institution.

It is difficult to imagine how to craft an effective statutory limitation on preemptive archiving of Web content that is based on either the characteristics of the content or the intended audience of the website. The Web changes too rapidly (one of its acknowledged strengths) to be amenable of this kind of classification. Even if such a limitation were desirable in theory, it would likely not be possible to devise an algorithm that could reliably implement it in automated harvesting software, thus placing an impossible burden of manual review on institutions.

Sites and Content

A website preservation exception should *not* be limited to a defined class of sites or online content. University libraries must be able to continue to collect materials that support their teaching, learning, and research. Also, it would be practically impossible to define a class of materials that sensibly falls outside of "cultural heritage." Finally, our best definitions to date already place an undue burden on libraries wishing to build their collections. For example, a domain name test fails in the case of a government agency that outsources the provision of tax-funded public information to a vendor operating from a "dot-com" site (e.g., the Web address for the U.S. Postal Service is http://www.usps.com/).

Restrictions and User Registration

It should be noted that "access restrictions" and "user registration" are not the same thing. Our goal is to respect "access restrictions" that prevent users from accessing restricted content. However, materials that are publicly available should not be restricted from collection — access restrictions such as user registration and certain kinds of service charge are not an indication that materials are either not freely available or restricted from archiving. For example, a site might simultaneously charge a membership fee to defray operating expenses while welcoming archival crawlers.

Opt-out Provision

An opt-out policy is desirable, but it is not necessary to have this in the code. A Web protocol should be developed that is used by the community. Both content owners and libraries need to have flexible opt-out policies that meet a wide range of needs for both. This may include the ability to:

- Remove only selected content from the archive, not all;
- Specify that the material may be archived, but not redistributed, or only redistributed to a limited group of users; and
- Request that access to the archived material be delayed for a specified period of time.

Prior Notification of Crawl

It is not necessary to notify site owners or operators ahead of the crawl. The burden should not be placed on libraries to contact site owners to collect content. This is an undue burden and will prevent collection activities. If content collection must be approved by site owners, valuable content will inevitably be lost. Often, the most critical information to capture is content generated by emergencies and catastrophic events. This information is fleeting, and is often gone before the author can even be identified. Further, the response rates for requests to capture Web content are very poor. In a situation like Hurricane Katrina, for example, content owners have more pressing concerns than responding to requests to archive their Web pages.

In addition, a site owner is not necessarily the content owner. Some sites, like blog sites and other multi-user sites, have no single owner. And, although most websites provide a contact for the webmaster, (a) the webmaster is likely not that owner, and (b) there is no guarantee that a human being is at the other end of the webmaster email address.

"No archive" Meta-tags

If content is publicly accessible, archiving institutions should be able to capture it. It should not be necessary to respect "no archive" meta-tags, robot.txt files, or similar technologies that block sites or pages from being crawled. Currently, robots.txt files and the like are not always used consistently or appropriately. For example, some sites that are clearly in the public domain (federal) have robots.txt files. Moreover, a robots.txt file ignores the issue of whether content may be archived (e.g., by a human being operating a web browser). The current robots.txt mechanism is crude, inexpressive, and often created without the knowledge or authorization of the rights holder.

Underlying Software

The library or archives should be permitted to copy and retain a copy of a site's underlying software solely for purposes of preserving the site's original experience.

Without “the underlying software,” content often proves to be unusable. Note that underlying software usually includes things such as javascript and stylesheets that are downloaded automatically by a web client during normal browsing, and without which a web page cannot be rendered.

Public Access and Labeling

Libraries must be able to make captured web content available to their community of users. For content that is clearly protected, access controls would be put in place to prevent unauthorized redistribution.

If the content is clearly in the public domain, there is no reason to require the lapse of a certain period of time before allowing access. In instances where a waiting period is appropriate, this could be included as part of an opt-out policy (e.g., “Please take my content down for now, but you can put it up in six months.”).

Labeling should be required to make clear that captured pages or content are copies preserved by the library or archives, and not from the actual site. Labeling of previously captured content is good practice. It helps the user, and it also maintains the integrity of the content (provenance) and the archive.

Patricia Cruse
Director, Digital Preservation Program
California Digital Library — Office of the President
University of California
415 20th Street
Oakland, CA 94612