Heritage Under Lock, but No Key:
The Troubled Status of Unpublished Works in Digital Archives Projects

William J. Maher, University of Illinois at Urbana-Champaign

October 14, 2005
Emory University Symposium on Free Culture and the Digital Library

Recent information technology developments, combined with increased attention to copyright law, have created a deep tension for archivists and librarians. While clearly interested in doing all we can to disseminate information, we also want to be good citizens and not tread on the rights of authors or publishers. No doubt, a system of creative commons licenses would allow greater dissemination of information, but these are of little value for those materials that fill archival repositories–pre-existing works locked in copyright but lacking any clear owners who might be approached for permissions. Indeed, the most important intellectual property challenge facing archivists and their users, as well as the one most in need of a legislative solution, is that of these so-called "orphan works." When neither scholars nor repositories can get permission to use these materials, public access is restricted, and both the archival mission and society suffer.

The 1976 copyright law was supposed to have clarified the murky territory in which archivists and their users had to work. The law ended the regime of perpetual copyright for unpublished works, but it has been of only limited use in supporting scholarly and public use. For example, because the fair use exemption is a weak, confusing, and costly device for the unpublished works, archivists receive frequent requests from authors whose publishers require written sign-offs for the use of a single quote or photograph. All too often, our response must be "We do not own the copyright and have no idea of who does." What's more, archival materials are overlooked by reform efforts, as seen in some responses to the Copyright Office's recent call for public comments on orphan works.

So what are diligent archivists to do when we receive the inevitable mandate to create a digital archives of unique items otherwise inaccessible outside the archives? Funding agencies and institutional legal counsels want clear-cut certification of copyright ownership or assignment, but what if, as is usually the case, the creators are long dead and their heirs virtually untraceable?

At the center of copyright law is Section 106 which provides that no one but the author or his/her assigns can reproduce or make copies, prepare derivative works, distribute copies, perform the work, or display the work. These are very substantial limits on what we can legally digitize and present online. Of course, the law provides two major limits that are seemingly relevant to digital libraries and archives. Section 107 (Fair Use), added to the law in 1976, is supposed to mediate between private and public interest. This is a laudable principle, but in practice the archivist encounters enormous difficulties in relying on fair use to build robust digital archives. Because fair use is an affirmative defense against a claim of infringement, it must be decided on a work-by-work and case-by-case basis in the course of a legal proceeding against the user. Fair use rules are not clear–they are overlapping and highly circumstantial–and certain kinds of transformative uses of archives have received little support in fair use decisions.

Turning to Section 108, one would hope that a section entitled "Limitations on Exclusive Rights: Reproduction by Libraries and Archives" might assist those building digital archives, but in fact Section 108 is narrowly construed to allow only limited copying for preservation purposes or one-time copying for end-users. Indeed, while the law now allows digital as well as analog copies, Section 108 does not allow such digital copies to be made available beyond the premises of the archives or library itself, a restriction inimical to the entire notion of the internet. Other

absurdities include Section 108 (h), which allows preservation copying and distribution of works in the last 20 years of their term, but applies only to published works; and Section 108 (i), which eliminates most of §108's exemptions if the work is audiovisual, musical, or pictorial some of the most sought-after archival materials.

A final issue is the length of term of copyright protection. Since 1978, all works, whether published or unpublished, are covered by copyright from the moment they are created. Thanks to 1998 entertainment industry lobbying, the term for all works is life of the author plus 70 years. For archives this presents several significant problems. First, few of the authors of works we hold are of such significance that their date of death can be determined, and in fact, given the commonality of names among the millions and millions of document creators, it is very often difficult to establish the identity of many of the authors, let alone locate all their heirs. Second, because most of the documents so valuable to archives are created as accidents of some other action, few authors leave means for the administration of their rights. Finally, the law's provision (§302 (e)) regarding presumption of death of the author does not really open up the possibility of using the works of untraceable authors until the documents are at least 120 years old, and then only with a cumbersome process of checking with the Copyright Office. If one's efforts to create a digital archives are thus limited to only such works as are clearly past their copyright term, virtually all of the history of the twentieth century has been fenced off from use.

So why not just focus on pre-twentieth-century items and ignore the rest? For archivists, that would contradict our core mission to be purveyors of recorded knowledge and thereby ensure that the knowledge created and accumulated by past generations is joined with that of the present, in order to make it available for society's future. Because knowledge is cumulative, and

because our work must result in an ultimate utility, we know that archives must be copied, quoted, published, and otherwise disseminated using the latest technology.

All archives, whether in government, educational institutions, professional associations, businesses, or churches, have not only a common mission but also base their work on some core archival principles. In fact, it is the nature of archival theory and methodology that makes the creation of a truly authentic digital "archive" in today's copyright world well-nigh impossible.

Why? First, the material that finds its way to archives is highly diverse both in physical format and in authorship. Any given correspondence file may contain hundreds, thousands, or even millions of separate copyrighted works and an equal number of authors, most of whom had little or no idea that their "works" would be deposited in a public repository, let alone might be disseminated through some "digital archives" project. Second, the very characteristic that makes archives so valuable as historical evidence--their spontaneous, almost accidental, creation–also means that few archival works have the artistic core or commercial utility that are the basis of American copyright law. Letters, photographs, sound-recordings, and other documents are, more often than not, created as the accidents of some other action, rather than as the a conscious creative expression. A third, and particularly important characteristic is that archives have a comprehensiveness that, while not absolute, is not compromised by artificial curatorial decisions. Indeed, the supposition that an archives is complete makes the documents it contains invaluable for constructing an accurate historical record so that readers can draw their own conclusions.

In a conventional environment, all we need to accomplish these archival goals, is attention to the archival principles of provenance, aggregate description, preservation, and equal

access. Accomplishing this same goals in a digital environment is significantly more complex. New technology enables the delivery of archival content globally without the costs of building a distribution network, but unlike the conventional world, the digital environment runs afoul of intellectual property law, even if the technical and resource issues can be resolved. That is because the items are no longer just held and examined under the "first sale" rights, but are also copied and distributed by display on a network.

Given the complex nature of archival repositories and given copyright's broad sweep and egregiously long terms, it is no wonder that there is such a limited number of robust archival digitization projects. This assessment may seem at odds with the appearance of several online historical projects with the word "archives" in their name, including those hosted by archives and manuscript repositories. In fact, on examination, one can see that the scope and depth of these efforts are severely limited. While the materials that have been mounted are clearly of use by themselves, the fact that a complete record cannot be displayed means these sites do not constitute a genuine or significant archival presence.

To assess how various institutions and consortia have dealt with this dilemma, I examined thirteen digital archives sites.[1] Since most of these projects function as portals to large consortia of institutions, and each institution generally has multiple collections and web products presenting "digital archives," there is a much larger number of collections and projects represented through these sites, and it should come as no surprise that it is virtually impossible to follow all links, quantify the results, or even apply uniform data collection tools. Thus, the following observations are summary but still instructive.

Scope and range of projects: There is little consistency among projects in terms of

content and depth of coverage. Some projects are heavily focused on providing consolidated, searchable EAD finding aids. Others put a primary emphasis on training, policy, and best practices. Yet others provide consortium-wide search engines for the subject content of collections, while still others settle for providing links to the home pages of the participating repositories, each with its own search tools.

Content of Digital Collections: Clearly the web presentation of finding aids is nothing new, but what has developed more recently is the first glimmerings of a more robust exploitation of the possibilities of the Internet. However, when one looks at the actual archival content of the various sites, it is obvious that there is a long way to go. The most common content seems to be individual photographs, generally selected from much larger collections. Another major component is clear public domain works–ones published before 1923 or U.S. government works, for example, the Making of America Digital Library. In far too few instances are there primary sources or unpublished textual documents, and these are almost always just isolated items selected from larger collections. With the exception of some largely pre-1923 items on LC's American Memory site, there is very little in the way of sound or audio-visual material.

Range of Institutions Represented: One striking characteristic is the extent to which public libraries and local historical societies have participated, a notable accomplishment since such repositories have traditionally not been party to large surveys of archives and manuscript collections. Although it varies by state or region, academic libraries and archives have participated, but not always to a very great extent (e.g., the Illinois Digital Archives Program). State archives and other governmental records repositories are under-represented, and almost totally absent are business and corporate archives.

Project and Consortia Intellectual Property Policy: Regarding copyright policy, the greatest consistency is at the consortia level, where most are emphatically aimed at protection against claims of contributory infringement. The projects generally state that the only materials to be digitized are ones in the public domain or those for which participating institutions have obtained written permission. They also normally contain a clause to "kick-out" any institution if it violates copyright. Unfortunately, consortial guidelines generally are silent on the problem of orphan works, although they often create the seeds of a justification for applying fair use when the owner cannot be located. The consortial guidelines are good at instructing web visitors of the limits on what they may do with the material. Overall, the policies provide clear evidence of having consulted educationally-based, though risk-adverse legal counsel.

Individual Institutions' Copyright Policy and Practice: On a policy level, most of the individual institution sites affirm adherence to the same copyright policies as the consortia. However, when one starts looking closely at the content, there is much variability. The following practices are representative of the lack of uniformity.

1. A number of institutions appear to want to play it safe and limit themselves to items published before 1923, items published by the U.S. government or by their parent institution (e.g., Illinois Wesleyan *Argus*), or to unpublished works written more than 120 years ago (e.g. *Valley of the Shadow*). This is fine if one's historical interests do not extend later than when there was a chance that the League of Nations could end conflicts among nations. A lot has happened however in the last eight decades of the twentieth century, and given the need to connect history to a generation that soon will not be able to remember any president other than "W," it is clear that such digitization projects will be of limited value in fulfilling their missions.

2. Some institutions present published and unpublished primary source items for which they have a plausible claim of ownership and present them in some depth. In a few instances, because the repository has negotiated a copyright transfer in the deed of gift, it is able to make important unpublished or even recently published material available–for example, Colorado State University's extensive wild animal photographs or the University of California, Davis' mid-century commercial photographs of California. Unfortunately, given the vast quantities of valuable historical archives donated over the years without copyright transfers and given the extent to which third-party correspondence and other works make historical collections valuable, this approach has major limitations.

3. Some sites claim ownership, even when the claim seems implausible according to information provided on the site. In some cases, they may be claiming nothing more than copyright ownership over the digitized image rather than of the underlying work, or they may be seeking to limit what can be done with the image because they are the sole-source provider of the image but this is a disservice to the educational purposes of digital archives.

4. Many institutions follow a "throw up their hands" approach. They indicate that they have made efforts to contact copyright owners without success but post the item(s) on the basis of fair use while also including a notice that should any copyright owner or other party have information about ownership, they should come forward. To show good faith, they sometimes include clauses indicating their desire to hear from any copyright owners, and in some further cases, they even promise they will remove material if a copyright owner does appear.[2] Overall, fair use of this sort seems a reasonable approach, although one that ultimately places you at risk of having to invoke the always murky four-factor defense. If followed faithfully, it also requires

considerable effort to document one's efforts at pursuing owners.

5. On many sites, the content presented is less than complete, and thus far from archival. In some cases, such as sites that include pre-1923 published material from a collection but no correspondence from relevant individuals, the reason is probably copyright. A prime example of incompleteness is the Paul Eliot Green Papers at the University of North Carolina where all that has been digitized is about sixteen letters, comprising 111 pages of 1917-19 correspondence from a 192-linear foot, 110,000-item collection. There may have been sound intellectual reasons for the choices made with the Green Papers, but I suspect this is not always true. In a case I know more intimately, the James B. Reston Papers at the University of Illinois, we have been able to digitized only about 2,000 pages of an estimated 146,000 pages in the collection, and the main reason for the limit is copyright. When individual documents are "cherry-picked" out of a complete collection, the project may be a digital scrapbook, a digital exhibit, but hardly a digital archives.

6. Ultimately, we cannot say that copyright barriers are the only reason for such selection and narrowing of the digital content. In some cases, there seem to be understandable cost and project management reasons for selectivity. For example, Colorado State University's wildlife photos project included only 1,000 out of a total collection of 20,000 images. These are not indefensible editorial decisions, but the result comes up short of the high hopes and promises of several digital library and archival projects.

Indeed all of these practices are quite at odds with the rhetoric by which such projects are promoted. For example, we are told that a driving concept for the Valley of the Shadow project was that it be "'a research library in a box, enabling students at places without a large archive

[sic] to do the same kind of research as a professional historian.'" The Online Archive of California goal of providing "all" with " . . . access to information previously available only to scholars who traveled to collection sites." is clearly undermined by the occurrence, at least 85 times, of the following line in bibliographic records for collections in the OAC database: "Items Online: None online. Must visit contributing institution."[3]

As this review makes clear, creating a true digital archives will always run afoul of copyright unless we can solve the orphan works problem for unpublished material. The issue at hand for this Symposium is where to go from here.

First, consortial projects and individual repositories must make clear that participants need to examine copyright ownership before digitizing and mounting materials on the web. The rule of digitizing only that which you own or for which you have permission is probably the only safe one, but it is the repositories which are in the best position to know the facts of a specific collection to determine whether a hard line or some is most appropriate.

Second, although fair use, as Lawrence Lessig has said, is not much more than a licence to hire a lawyer, projects should look to it to establish the context for digitizing and displaying material for which copyright owners cannot be readily located and which can be justified for their educational, non-commercial, cultural value. There is no assurance of protection from litigation, but if a repository's investigation shows no existing market for the works, and if the site includes appropriate disclaimers, then fair use is the only present basis for digitizing the orphaned copyright works that must be included if there are to be meaningful digital archives.[4]

Third, the library, archival, and internet community should make a focused effort to amend Section 108 (h) of the copyright law so it includes unpublished works in the scope of

materials that libraries and archives can digitize and make accessible in the last 20 years of their copyright term. Better still would be adoption of a full-scale orphaned works exemption along the lines supported by the Society of American Archivists.[5] Archives, far more than published works, are very likley to be orphaned material, often for the same reasons that they are valuable research materials: they contains a multiplicity of authors, those authors are virtually anonymous, it is unclear those authors ever constructed their works for dissemination, and the works themselves are of research value but almost always of limited or no commercial value.

Public policy efforts should also focus on the international level to ensure a solid basis for the kind of safe haven needed to allow digital library projects to include sufficient archival material to make them credible digital archives. Specifically, we need to see support for allowance within the Berne/WIPO treaties for non-commercial, educational use of unpublished works. Such efforts have been initiated in the International Federation of Library Associations (IFLA) and the International Council on Archives (ICA) following discussions at their separate 2004 congresses. However, funding is needed to bring together the librarians, archivists, and international copyright law specialists to craft language to be advanced to WIPO. Then, a concerted effort will be necessary to have this kind of change adopted.

In conclusion, the many digitization projects to date have made a noble effort to expand the public's access to cultural research materials beyond those previously at hand through local libraries, but they can hardly be called "archives" in the full sense of the term because they have been unable to provide very deep or broad access to much truly archival material. In many instances, quite understandable cost and pragmatic hurdles have caused these efforts to be rather limited. However, the significance of the copyright law barrier is hard to overestimate. For any

meaningful, robust, on-line digital archives to exist, the copyright issues must be addressed and the barriers they create reduced to manageable hurdles at the most. We need sophisticated keys that allow us to unlock these works. Archivists and librarians should not become gatekeepers locking away orphan works thus serving neither the original author, the works, or the archives-using public.

REFERENCES

1. They were: Colorado Digitization Program, Northwest Digital Archives, Northwest Digital Archives, North Carolina Echo (Exploring Cultural Heritage Online), Making of America, Cornell University Library Digital Collections, New York University's "The Database of Recorded American Music, Library of Congress' American Memory Project, University of Virginia's Valley of the Shadow Project, American Museum of Natural History, Tufts University's Perseus Project, the Illinois State Library's Illinois Digital Archives, and the University of Illinois' American Library Association Archives Digital Collections.

2. A balanced statement is that from Cornell University: "The Kheel Center would like to learn more about these images and hear from any copyright owners who are not properly identified on this Web site so that we may make the necessary corrections." They then provide a staff name and e-mail address. www.laborphotos.cornell.edu/copyright.php?Kheel=7e23269dd4c420e8c06ea581a1f9e73e

3. At more detail level, in one instance of a 43 linear foot photographic collection, 54 images have been digitized, and when one looks more closely, those images come from only 45 percent of the folders in one subseries of that collection.

4. For example, that from the American Museum of Natural History site reads: "While this Website is publicly-accessible, not all the materials are in the public domain -- the majority of the images, texts and data are copyrighted to the American Museum of Natural History -- and a number of other texts and images are still copyrighted to their original print publishers or digitizers and made available here with permission. We have put great effort and expense into producing this site, and we hope the results are useful to a broad audience." http://library.amnh.org/diglib/conditions.html

5. The SAA initial comment is at: http://www.copyright.gov/orphan/comments/OW0620-SAA.pdf, and the "reply comment" is at: http://www.copyright.gov/orphan/comments/reply/OWR0088-SAA.pdf