Only with Your Permission: How Rights Holders Respond (or Don't Respond) to Requests to Display Archival Materials Online

[Final publication is available at http://www.springerlink.com/content/u4g3847818t61w84/]

Dharma Akmon
School of Information
University of Michigan
1075 Beal Avenue
Ann Arbor, Michigan 48109-2112

dharmrae@umich.edu
734-395-0790

As patrons increasingly expect remote access to information resources and archival institutions recognize the possibilities that online access enables (e.g., full-text search, less handling of fragile materials), more repositories are attempting to digitize collections in their entirety so that researchers can access them through publically accessible web sites. While there are plenty of challenges to such projects (e.g., costs, time, interface design), copyright represents a significant and sometimes insurmountable obstacle.

Because of complicated rules, a lack of case law, and differences based on where in the world a work was created, copyright law is a noted area of confusion to archivists (Dryden 2008). It is difficult to describe copyright protection in a succinct way, and it is not the purpose of this paper to go into detail about the law. A number of resources cover U.S. copyright law extensively, specifically addressing the concerns of libraries and archival institutions; of particular note are the resources provided by Stanford University's Copyright and Fair Use Center (Stanford University Libraries 2007). Most relevant to the current study is that U.S. copyright law grants exclusive, limited rights to copyright holders to reproduce, publish, and publically display their works. In the United States, copyright is automatic for original works created on or after January 1, 1978, meaning that registering a work with the Copyright Office and affixing a copyright notice are not prerequisites to protection. The length of copyright protection depends on a number of factors, including the date the work was created or published, the date of the death of the rights holder, whether any transfers were made, and whether registration was renewed, making determining copyright status even more difficult (Stanford University Libraries 2007). Although the research presented here was carried out in the context of U.S. copyright law, archivists everywhere face some variation of this problem.

While the transfer of materials to archival repositories usually involves a formal deed of gift that specifies access to the collection and whether or not the donor retains copyright, the donor frequently does not have the legal right to specify copyright terms for the entire collection. She might have owned the physical collection, but if it contains documents authored by third

parties she does not hold the copyrights to those items unless the third party authors or publishers transferred those rights. As a result, the archival repository might not have the right to duplicate and distribute those materials online without permission from each copyright holder represented in the collection.

Archival institutions employ several different strategies to respond to the barrier that copyright law represents to online, open access: they avoid digitizing collections with complicated rights issues (e.g. not in the public domain, copyright not held by the archives, or copyright held by a third party); they interpret their digitization and distribution of materials online as "fair use", or they attempt to identify, locate, and seek permission for every item they plan to put online. Each of these options has obvious drawbacks. In the first instance, archival repositories limit themselves to a relatively narrow set of collections or to digitizing only parts of collections. In the second option, they risk litigation from rights holders since fair use is not an explicit set of rules but, rather, a set of guidelines applied by the courts on a case-by-case basis. And in the last instance, they must invest significant resources in a task whose outcome is uncertain and which may represent too strict of an interpretation of the law (e.g. they are not exercising fair use at all). Additionally, the archival literature offers little information on what effort is actually required to seek permission from rights holders and what the results of those efforts might be.

This paper presents an analysis of data on copyright clearance collected from the Jon Cohen AIDS Research Collection digitization project at the University of Michigan Library and the School of Information, which took place from 2007-2009. The Cohen collection contains 13,381 items, 5,463 (approximately 11 linear feet) of which are protected by copyright held by 1,377 unique copyright holders. Because of the Library's high aversion to the possibility of copyright litigation, project staff were required to obtain permission for each copyrighted item prior to digitization. In this paper, I analyze data gathered during the copyright permissions process to answer the research question: which copyright-related factors should archives

consider when evaluating which collections to digitize for online, open access? To get at this question, I will specifically investigate the following questions:

- How much effort is required to obtain permission from rights holders to place archival materials online?
- Is there a length of time after which archival repositories are unlikely to get a response from rights holders?
- How do rights holders respond to copyright requests?
- Are certain types of copyright holders more likely to deny permission than others?
- Are there characteristics of documents that are predictive of denial?

Answers to these questions will help archival institutions define appropriate strategies for dealing with the copyright issues associated with mass digitization projects. Additionally, this study presents data that will help archives predict the difficulty of attaining rights for the online display of collections.

Literature Review

While there are few empirical studies to guide archival repositories in copyright matters, there have been many efforts to describe the implications copyright law has on digitization projects. Work in this vein tends to be based on an interpretation of copyright law that leads the authors to conclude that archival repositories must only post materials that fall into one of three categories: in the public domain, copyright held by the archives, or materials for which the archives has been granted permission to display. Dismissing fair use as a viable option, experts argue that the copyright status of the items in the collection be a basis for selection decisions and that institutions should not digitize collections that they do not think they can *easily* obtain rights to (Hughes 2004; Lee 2001; Sitts 2000). However, none of these authors offers a model for predicting how easy or difficult it will be to obtain rights. Nor do they provide, as a basis of comparison, any benchmarks defining a reasonable effort. Instead, it seems that collections that fall into the “easy” category are those that are old enough to be in the public domain, those

for which the repository owns copyright already, or those which consist of documents from a single copyright holder. All others are "difficult" and should not be digitized for online access.

Others (Besek 2003; Minnow 2002), while recognizing fair use as an option for digitization projects, admit that the fact that "there is no magic formula to determine whether a use is fair" poses a risk to archival repositories that wish to exercise this exemption (Besek 2003, p. 6). However, as both Besek (2003) and Hirtle (2001) note, some repositories are posting works that they have not gained rights to with a plan to remove the documents on request from the rights holder.[1] This strategy may be particularly important for archival repositories engaged in digitization projects since it can be especially difficult to identify and locate the copyright holders of unpublished documents (Hirtle 2001).

Hirtle (2001), calling copyright status investigation "an unacceptable tax on the scholarly research process," recommends that those who plan to make unpublished materials available to the public should distinguish between materials created with no intention for profit and those which were intended for sale or created by those who made a living by their writing. Additionally, he argues that archivists and scholars must weigh the benefits and risks to both scholarship and rights holders to determine whether or not they should make the items available to the public.

As for how archival repositories are actually responding to copyright issues, Dryden (2008) provides the most extensive study to date. Her study focused on Canadian repositories with online holdings to determine how copyright law has influenced their selection decisions and how they go about getting permission to post copyrighted documents. Her interview and questionnaire data suggest that archivists do consider copyright issues when selecting materials to put online. They assess risk (on the basis of material type and date, for example) and include or do not include items based on their personal tolerance for the perceived risk and their

[1] An example of such a statement can be found at the Library of Congress American Memory Projects. Retrieved from http://memory.loc.gov/ammem/coolhtml/ccres.html, on October 5, 2009.

impression of their institution's tolerance for risk. Overall, Dryden found that archivists prefer to select items for which their institutions already hold copyright or that are in the public domain, and that they are generally "reluctant to use material in which they have not obtained copyright clearances" (Dryden 2008, p. 136).

Dryden also found that when repositories seek permission to display items, they go through significant effort to do so, tracking down identities of rights holders, their current locations, and contacting them via various means (phone, email, and post). Two-thirds of Dryden's questionnaire respondents indicated that if they were unsuccessful in locating a copyright holder or did not get a response, they would not use the document, perhaps substituting an acceptable replacement for it. The few who said they would post documents online anyway said they would only do so if the material was not risky (for example, sound recordings were considered by many to be a risky type of material) or if there was not likely to be someone with a financial interest in the material.

There are few studies that report the outcome of copyright efforts for digitization projects in archival repositories, and, for the most part, the data reported are limited. Some digitization project reports focus on copyright issues specifically, but data on the amount of effort invested and analyses of success rates are missing. Cave, Deegan, and Heinink (2000) report on their experiences with the copyright process at the Refugee Studies Centre Digital Library project at the University of Oxford. This repository consists of primarily modern (e.g. still in copyright) 'grey literature'. They report that the process of seeking permissions took a significant amount of resources even though they did not actually pay any royalties for rights. It cost approximately £5-6 per document the first year (when they were creating and refining their process) and then £2-3 per document after the first year. An important factor in the cost per document was whether the document was the only one for a copyright holder or one of several; it was significantly more expensive to seek rights for a document when it was the only one held by a particular rights holder. Unfortunately, while the authors report that it took a considerable

amount of time just to identify and locate copyright holders, they do not provide time data for tasks individually or as a whole.

In her article about the Ad*Access project at Duke University, Pritcher (2000) provides insight on the effort to digitize and put online over 7,000 early- to mid-19th-century U.S. newspaper and magazine ads. Pritcher primarily focuses on how the Library determined whether or not its project fell within fair use and what the implications of that determination were for the project. Project staff decided that, while Duke is an educational institution and the collection was to be used for research (two factors that might favor a fair use argument), online access to the items raised new issues that had not been addressed in a clear way by the courts. As a result, they ultimately decided to seek permission for the items.

While they were fortunate that their rights holders were companies instead of individuals (who might be harder to locate), project staff still had to contend with numerous mergers and dissolutions that challenged their efforts to identify and contact the companies. Because of difficulties in identifying and locating rights holders, they amended their selection process later in the project to include only those items from companies with a large number of ads in the collection or that were "well-known, recognizable names." Pritcher notes that a significant amount of time was devoted to re-contacting individuals who did not respond to initial requests, but does not report how much time except to say there could often be a six-month lag between initial contact and resolution. As with the Refugee Studies Centre Digital Library project, Pritcher offers only a rough estimate of cost (without time reports) that comes out to approximately $1.43 per advertisement.

The most complete set of data provided on copyright efforts associated with mass digitization come from Covey (2005) and George (2002, 2005) on a set of projects at Carnegie Mellon University Library to seek permission to offer online access to primarily published, out-of-print books. In the Random Feasibility Study conducted 1999-2000, they selected a random sample of titles from their library. They sent letters to the copyright holders represented by

these items, requesting permission to include them online. One of the first lessons they learned was that the process is “time-consuming and often unsuccessful" (Covey 2005, p. 13). They were unable to find addresses for 7% of the publishers, and many of their letters were returned ‘address unknown’. Ultimately, they were unable to locate 21% of the copyright holders, and well over half of those they did locate required second and sometimes third letters. Of the publishers who responded to their requests, more than half granted permission, but 68% of the grantees stipulated some kind of access restriction.

Covey and George did not keep track of transaction costs during the study, but estimate it cost approximately $200 per title that was granted. They conclude that, while doable, identifying and locating copyright holders and then negotiating with them requires a significant investment of resources and that the investment is often met with non-response. This conclusion resulted in a different approach for their Million Books Project. Instead of seeking permission on a per title basis, project staff sought a per publisher copyright permission model that would allow the Library to place all titles by a publisher in the collection without having to list all of those items upfront for the copyright holder. They were able to bring costs down to 69 cents per title using this model and conclude it is the only one that scales up. Unfortunately, this is probably not a strategy that archival repositories could employ successfully, since they generally hold such a large amount of unpublished materials for which individuals and organizations, not publishers, hold copyright. To request blanket permission for any document an individual or organization might have created is far different from requesting permission for all published items from a publisher.

Interest is growing in the ways that copyright affects digitization efforts in archival repositories, but it is evident that the field is lacking in studies of the effects of copyright on archival repositories’ digitization projects and the outcomes of efforts to attain permission to put materials online. While it is clear that seeking copyright permission takes a significant amount of time and resources, we do not have data that indicate how much. Further, no one has used

data to analyze possible predictors of denial and acceptance so that repositories might have more precise guidance in deciding which collections to digitize. While the projects at Carnegie Mellon resulted in extensive data about the process of seeking permission for published, out-of-print books, we do not yet have a similar study of archival materials. It is the aim of this paper to fill that gap.

Methods

Data for this study were gathered from database records kept during the copyright permission process for the Jon Cohen AIDS Research Collection Digitization Project.[2] Jon Cohen, noted writer for *Science Magazine* and author of *Shots in the Dark: The Wayward Search for an AIDS Vaccine*, donated his collection of AIDS-related material that was used as research for his writing. The collection arrived at the Library unprocessed, with staff uncertain as to how many copyright holders would be associated with the collection and what portion of it belonged (from a copyright standpoint) to Jon Cohen. Once processed and inventoried, the collection was comprised of 13,381 items (see Figure 1 for detail). Of these, 6,026 (45%) items were newspaper and journal articles, which project staff decided they would not seek permissions to digitize. Since these articles are generally offered through fee-based databases, the online collection instead presents metadata and SFX links, an OpenURL link resolver that utilizes article or book metadata to link to the full-text version of the item, for each article to search users' libraries. Of the non-article items, 1,892 (26%) were U.S. Federal government documents, which (because they are automatically in the public domain) could be digitized and offered openly online without permission (CENDI 2008). This left 5,463 documents in the collection that were in copyright, of which only 209 (4%) were copyright held by the donor, Jon Cohen. Cohen Project staff had to identify, locate, and obtain permission for all copyrighted items before including them in the online collection. At the peak of the project, between May of

[2] The Jon Cohen AIDS Research Papers can be found at http://quod.lib.umich.edu/c/cohenaids/. Retrieved October 28, 2009.

2007 and August of 2008, there were at all times at least two, and sometimes as many as four, staff people working between 15 and 30 hours a week on Cohen Project tasks.

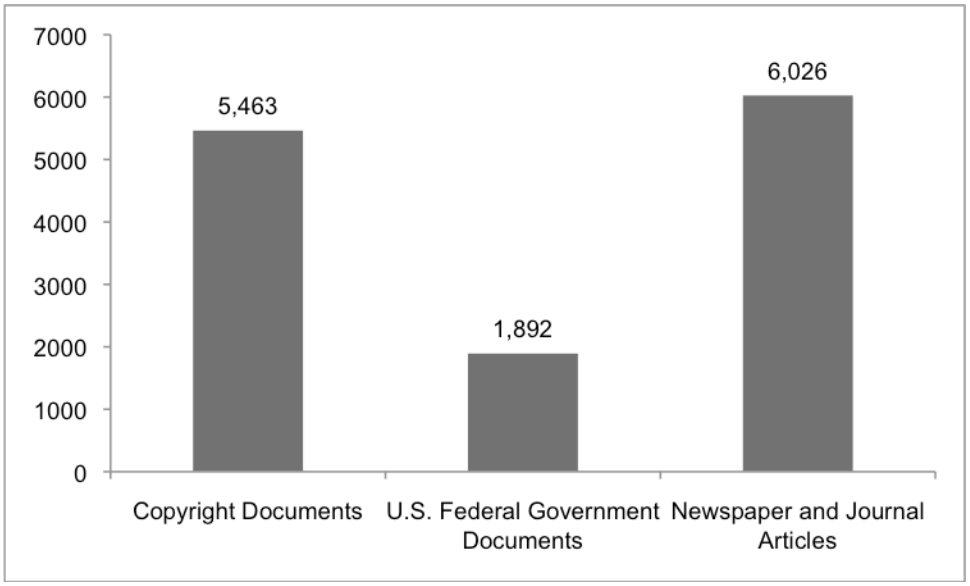

Figure 1: Composition of the Cohen Collection (N=13,381). This paper deals primarily with Copyright Documents.

Copyright experts at the Library urged project staff to keep track of each collection item in a database along with a record for each rights holder so that they could more easily carry out the permissions process and have documentation of their efforts. In this database, project staff recorded item-level metadata, including a unique ID for each item, title (staff-created if not provided by the document), creator name(s), creation or published date, genre type (based on the Getty Art & Architecture Thesaurus), copyright holder name(s), and permissions status. Each item record was linked to its associated copyright holder record(s). Each copyright holder record contained contact information, a record of communications, and the final outcome of permissions requests.

Including Jon Cohen, project staff identified 1,377 unique copyright holders. They sent each copyright holder for which they could find contact information a letter describing the project and how the materials would be used, and requesting non-exclusive rights to include their material (listed item by item on an attached page) in the online collection. The Library did not

offer to pay fees for this right and was asked by just one rights holder to pay a royalty fee (the Library refused, and the rights holder subsequently granted permission). Permissions requests also included letters of support from Jon Cohen and John D. Evans, founder of the project's funding agency, the John D. Evans Foundation. Copyright holders were instructed to sign and date a consent form, signifying their agreement to the terms and were provided with an addressed envelope in which they could return the form. Rights holders with North American addresses could return letters in the enclosed business reply envelope, which charged the postage to the Library. All rights holders also had the option of returning the form via fax or email. Finally, rights holders were provided with staff contact information in case they had further questions.

The Library considered non-responses to be equivalent to denials in that only items with explicit approval could be included in the online collection. For that reason, and because there was a dedicated source of funding for the project, staff made considerable effort to locate, contact, and obtain definitive responses from rights holders, often attempting to contact them multiple times and by various means (postal mail was used first when possible, often followed up with email and/or phone calls, depending on the information available).

Some of the copyrighted items (201 of 5,463) in the collection had more than one copyright holder associated with them. Library policy was to attempt to find contact information for and obtain rights from each copyright holder for a given item. However, consent from at least one copyright holder (assuming there were no denials) was enough for staff to consider an item safe for inclusion in the online collection. If any of the rights holders for an item with multiple rights holders denied permission, staff considered the item denied for inclusion in the online archival collection.

While staff did not include copies of rights holders' documents with the permissions requests, many copyright holders contacted the Library, requesting copies of the documents

listed in the permissions letter. In most instances staff responded by emailing or faxing copies of the documents in question.

It should be noted that, because staff were unsure of what kind of response to expect, they began sending out permissions letters before the entire collection had been cataloged. As a result, 83 (8%) of the copyright holders that staff contacted received two separate requests, listing two different sets of items: those that had been cataloged at the time of the first letter, and those that had been found after completely cataloging the collection. In much of the following analysis, these are treated as separate, independent requests.

Staff also kept track of time spent doing various project tasks. Time (to the nearest quarter hour increment) was recorded daily by each of the project staff as belonging to one of the following categories: 1.) processing and arrangement, 2.) encoding EAD, 3.) preparing items for digitization, 4.) entering item-level metadata (including associated copyright holder records) into the database, 5.) gathering contact information for copyright holders, or 6.) contacting copyright holders. The actual digitization was not included in labor calculations for two main reasons: it was done by an outside vendor making it infeasible to attain accurate measures of the time spent imaging the Cohen collection; and there are numerous guides that focus on imaging costs, but few that focus, as the current study does, on permissions tasks.

At the conclusion of the project, to aid in data analysis, a field was added to the copyright holder records to note the type of copyright holder. The following groups became apparent in the course of looking at the copyright holders represented in the Cohen Collection:

- Individual-- This category covers copyright holders as individuals, as opposed to organizational affiliates. An example would be John Q. Citizen for a personal letter he sent to a friend.
- Non-profit-- Organizational copyright holders that exist to provide programs and services of public benefit. Includes many hospitals, foundations, and non-governmental research institutes.

- Government (excludes U.S. Federal government since these documents are in the public domain)-- Governmental body copyright holders, including U.S. state, non-U.S. government, and intra-governmental organizations, such as the United Nations.
- Education-- Any educational institution copyright holders, including universities and their teaching hospitals and research institutes.
- Association-- Organizational copyright holders characterized by the affiliation of a group of individuals formed around a common interest or purpose. Examples in this collection were professional associations, societies, and many activist organizations.
- Commercial-- Organizational copyright holders engaged in commercial enterprise. Examples in this collection were pharmaceutical companies, corporate laboratories, and many publishing companies.

Results

The Composition of the Collection

Of the 5,463 copyrighted items in the collection, project staff were able to identify and find contact information for at least one rights holder or heir for 4,776 (87%) items. Of the remaining 687 (13%) items, 356 (52%) belonged to companies that had become defunct with no identifiable organization taking over copyrights; 284 (41%) belonged to rights holders for which staff could not find contact information; 25 (4%) belonged to copyright holders who are deceased with no known or located heirs; and 22 (3%) of the items did not contain enough information to identify any specific rights holder (See Table 1).

Table 1: Numbers of Copyrighted Documents by Staff's Ability to Request Permission

		Number of Items (Documents)	Totals
Able to Contact Rights Holder		4,776	4,776
Unable to Contact Rights Holder Because:	Company Defunct	356	
	Contact Information Not Available	284	
	Deceased with No Known Heirs	25	
	Could not Identify Rights Holder	22	687
			5,463

Of the 1,377 unique copyright holders represented in the collection, the largest portion (48%; 658) was made up of copyright holders in the "Individual" category. Table 2 shows the number and percent of each of the six different types of rights holders.

Table 2: Types of Copyright Holders in the Cohen Collection

	Number	Percent
Individual	658	48%
Non-Profit	220	16%
Commercial	196	14%
Government	99	7%
Education	90	6%
Association	65	5%
Unable to Categorize	49	4%
Total	1,377	100%

Project staff spent considerable effort gathering copyright holder contact information, primarily through online search engines. Though time intensive, staff were able to find contact information in the form of a mailing address, phone number, and/or email address for the majority, 74% (1,023), of the identified rights holders represented in the collection.

Effort Required to Request and Obtain Copyright Permission

The project time taken to accomplish copyright-related tasks eclipsed all other tasks tracked for the project. "Copyright-related tasks" are defined here as all tasks that were carried out solely because of rights issues. Specifically, that includes entering item-level metadata in the database (because staff did so only because of rights issues), gathering copyright holder information, and contacting copyright holders. For the Cohen Project, 85% of total project staff time was spent on copyright tasks (Figure 2 shows a detailed breakdown of each major type of task).

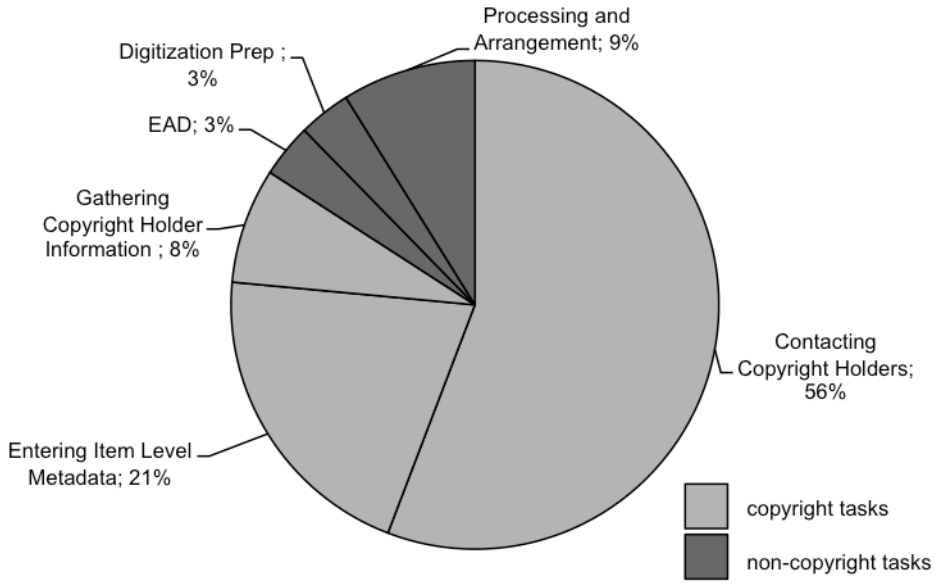

Figure 2: Relative Time Spent on Cohen Project Tasks

Some of the copyright-related work should be thought of on an average per document basis (entering in item-level metadata), while other work should be considered on a per copyright holder basis (gathering copyright holder information and contacting copyright holders). For the Cohen Collection it took staff, on average, 4.66 minutes per document to enter item level metadata and 70.3 minutes per rights holder to gather contact information and contact and negotiate with rights holders. Clearly, from a labor standpoint it is better to have a high item to

copyright holder ratio. That is to say, the Cohen data show that to get permission for twenty documents from one rights holder takes approximately 1.4 hours on average, but to get permission for twenty documents from twenty unique rights holders would take, on average, 12.5 hours.

While most of the copyrighted items (96%, 5,262) in the Cohen collection had just one copyright holder, a small portion had between two and ten copyright holders. Items with multiple rights holders would generally take considerably more time to obtain permissions for than items with just one rights holder. However, because consent from any one rights holder for an item was considered enough to put the item online (so long as there were no denials), staff often did not attempt to contact a rights holder more than once if that particular rights holder only held permission to documents that had already been granted permission by at least one of the other rights holders.

Copyright Holder Response

Project staff were able to determine the type of copyright holder for 744 of the answered permissions requests. The mean response time from staff's initial permissions request until resolution (where resolution is defined as some answer from the rights holder), was 41 days ($CI_{.95}$ = 38.49, 43.75). ANOVA was used to determine any differences in mean response times between the six different types of copyright holders. Levene's test of homogeneity of variance suggested that the six groups did not have equal variance in response time, so Welch's Robust test of differences in means was used in the one-way ANOVA. Welch's test suggests that there were significant differences in the means ($F_{Welch}(5, 171.7) = 7.68$, $p<0.001$). Specifically, when performing multiple comparisons of the means using Dunnett's T3 test (given the unequal variances), commercial copyright holders were found to have a significantly higher mean response time (57 days; $CI_{.95}$ = 48.15, 66.80) than individual copyright holders (33 days; $CI_{.95}$ = 33.03, 36.03). No other differences in mean times between types of copyright holders were found to be significant.

For this group of requests, there were also significant differences in mean response times for different types of answers. Possible responses to permissions requests in this comparison were: “refuse permission for all”, “grant permission for all”, and “grant permission for some” (where “all” means all documents listed in the request). Looking at only those requests for which there was one of these three responses, Levene’s test of homogeneity of variance again suggested that the three groups did not have equal variance in response times, so Welch’s Robust test of differences in means was used in the one-way ANOVA. Welch’s test suggested that there were significant differences in the means ($F_{Welch}(2, 40.70) = 14.15$, $p<0.001$). When performing multiple comparisons of the means using Dunnett’s T3 test, “refuse permission for all” responses were found to have a significantly higher mean response time (64 days; $CI_{.95}$ = 50.66, 77.70), than “grant permission to all” responses (38 days; $CI_{.95}$ = 35.63, 40.70). In addition, “grant permission for some” responses also had a significantly higher mean response time (83 days: $CI_{.95}$ = 59.05, 107.04) than “grant permission to all” responses. There was no significant difference in mean response times between “refuse permission for all” and “grant permission for some” responses.

Staff made 1,092 permissions requests to rights holders that could be classified into one of the six copyright holder type categories. To answer the question of whether or not there is a length of time after which archival repositories are not likely to get a response from rights holders, a Kaplan-Meier estimator was used to calculate the survival curves of each of the six types of copyright holders. Survival is defined as non-response, with the event being receiving a response from the rights holder (see Figure 3).

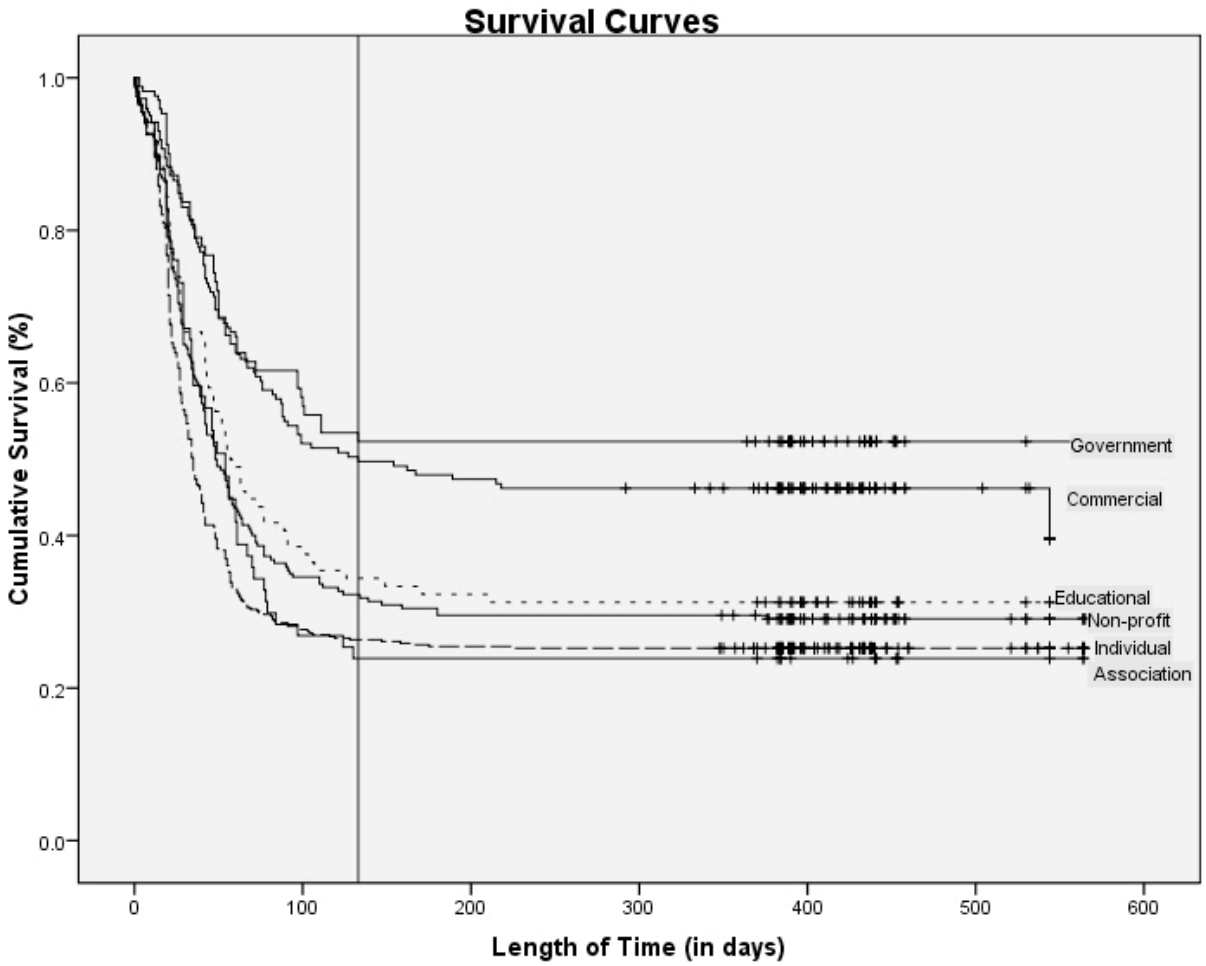

Figure 3: Survival Curves Based on Type of Copyright Holder (hash marks indicate the number of days after which rights holders were censored from the graph. It is possible that we may have heard from rights holders after this length of time had we still been making attempts).

These estimated curves suggest that approximately 120 days after the initial request, there are diminishing returns, no matter the type of copyright holder. That is, the percentage that survive (i.e. do not respond) stays level for all types of copyright holders after a period of approximately 120 days. A log-rank test was used to compare the survival curves for different types of copyright holders, and it suggested the curves were, in fact, significantly different from each other ($X^2(5, N = 1{,}092) = 58.6$, $p<0.001$). Specifically, a larger portion of the commercial and government copyright holders (about 50%) had not responded to the Library's requests at the 120-day mark than those in any of the other four categories.

Analyzing the responses staff received to requests, the good news is that most often those responses were to "grant permission to all." Of all of the responses from rights holders, 679 (91%) were to grant the Library permission to display all items requested in permissions letters. Only 46 (6%) of the responses were "refuse all." In fact, the biggest obstacle to displaying materials online was non-response: 352 (32%) of all of the Cohen Project permissions requests were met with non-response. Figures 4 and 5 depict the relative proportion of types of responses, first for all permission requests and second for only answered requests.

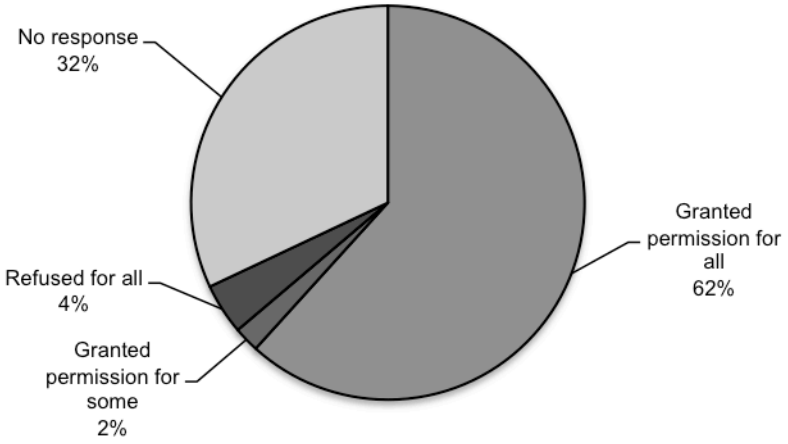

Figure 4: Responses to All Outgoing Copyright Permission Requests (N= 1,100)

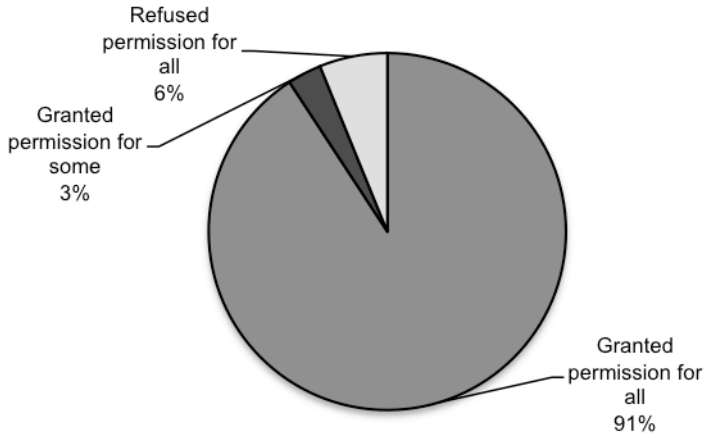

Figure 5: Responses to Answered Copyright Permissions Requests (N=748)

The result of this type of response is that the Library ultimately had permission to display online 3,490 (64%) of the copyrighted items in the collection. Unfortunately, 981 (18%) of the copyright items in the collection could not be displayed due to non-response from rights holders. Another 687 (13%) could not be displayed for three main reasons: staff could not identify the rights holder (22); staff could not locate the rights holder (309); or the rights holder was a company that they found to be defunct (356). Only 294 (5%) of the copyrighted items in the Cohen Collection were explicitly denied.

In the interest of determining whether certain types of copyright holders were more likely to "refuse permission for all" than other types, an overall chi-square test of association between copyright holder type and refusing all suggested that this association was significant (Pearson $X^2(5, N = 1{,}092) = 24.57$, $p<0.001$). Specifically, 19 (11%) of staff requests to commercial copyright holders led to a response of "refuse all" with the next highest 3 (5%) for association, which is close to the overall marginal percentage of "refuse all" responses (see Table 3 for detail).

Table 3: Responses to Requests for Different Types of Rights Holders

N=1,092	Refused All	Granted All	Mixed	Non-Response	Total
Individual	13 (2.9%)	320 (70.8%)	5 (1.1%)	114 (25.2%)	452
Non-profit	6 (2.7%)	146 (66.4%)	4 (1.8%)	64 (29.1%)	220
Government	3 (3.5%)	36 (41.9%)	2 (2.3%)	45 (52.3%)	86
Educational	2 (2.1%)	59 (61.5%)	5 (5.2%)	30 (31.3%)	96
Association	3 (4.5%)	48 (71.6%)	0 (0%)	16 (23.9%)	67
Commercial	19 (11.1%)	67 (39.2%)	7 (4.1%)	78 (45.6%)	171
Total	46 (4.2%)	676 (61.9%)	23 (2.1%)	347 (31.8%)	1,092

An overall chi-square test of the association between copyright type and "grant permission for all" responses suggested that this association was also significant (Pearson X^2=(5, N = 1,092) = 71.80, $p<0.001$). Almost 72% of requests to association copyright holders

and 71% of requests to individual copyright holders were met with an answer of 'grant permission for all'. Commercial copyright holders were the least likely to grant permission to all items in the request: only 39% did so, compared to the marginal rate of 62%. The rest of the data for this comparison can be found in Table 3.

As the survival curves suggest, and a chi-square test confirms, there was also a significant difference found between copyright holders in their rate of non-response (Pearson X^2 (5, N = 1,092) = 43.49, p<0.001). While the overall non-response rate was 32%, 52% of requests to government copyright holders were met with non-response, followed by 46% of commercial copyright holders. Table 3 shows the overall marginal percentage of non-response and comparison of other copyright type groups.

Characteristics of Documents That Predict Denial

The previous sections provided an analysis of the correlation between copyright holder type and "accept all," "deny all," or non-response to permissions requests. The Cohen Project data also reveal that there are traits of *documents* that are predictive of permissions denial and acceptance. Specifically, published status, year of creation or publication, and the type of copyright holder(s) associated with a document were all correlated with an item being denied permission. Staff considered "published" items to be those that appeared to have been to distributed to "the public by sale or other transfer of ownership, or by rental, lease, or lending," as has been defined by U.S. copyright law. For example, a press release would be considered published, while a person-to-person letter would be considered unpublished (United States Copyright Office 2009).

Looking at the 3,780 copyrighted items that were either granted or denied permission and which could be classified according to published status, an overall chi-square test of association between published status and whether an item was denied found a significant association between the two (Pearson X^2(1, N = 3,780) = 10.67, p = 0.001). Unpublished items were slightly more likely to be denied that published ones. Specifically, 9% of unpublished

items were denied, compared to 6% of published and the overall marginal percentage of denial at 8%.

To test the association between the creation date of items and denial status, items were categorized into one of five different date range categories representing the earliest and latest dates of items in the collection: 1941-1985, 1986-1990, 1991-1995, 1996-2000, or 2001-2005. Comparing the different groups, an overall chi-square test of association between date category and item refusal status suggested that the association was significant (Pearson X^2=(4, N = 3,365) = 10.37, p = 0.035). On closer examination, this association is driven primarily by items that fall into the 1991-1995 and 1996-2000 categories. Nine percent of the items that fell into the 1991-1995 category were denied, while only 6% in the 1996-2000 were denied (see Table 4). Belonging to one of the other date categories did not contribute significantly to the chi-square statistic.

Table 4: Item Denial Rates for Different Year Categories

N=3,365	Accepted	Denied	Total
1941-1985	36 (90.0%)	4 (10.0%)	40
1986-1990	261 (90.0%)	29 (10.0%)	290
1991-1995	1,399 (90.8%)	142 (9.2%)	1,541
1996-2000	1,090 (93.6%)	75 (6.4%)	1,165
2001-2005	309 (93.9%)	20 (6.1%)	329
Total	3,095 (92.0%)	270 (8.0%)	3,365

Lastly, the type of copyright holder(s) associated with the documents was also correlated with denial. Because a given item could have multiple copyright holders and, as a result, more than one type of copyright holder associated with it, each copyright type was examined separately. An overall chi-squared test was performed on each type and suggested a significant association between each type, except individual, and the denial status of an item. The most striking difference in denial rate was for items with at least one commercial copyright holder. Almost a quarter (24%) of these items were denied compared with 4% of items without

a commercial copyright holder. Table 5 lists the *p*-values for each test and Tables 6, 7, 8, 9, and 10 depict denial rates for each type of copyright holder where a significant difference was found.

Table 5: Association Between Type of Copyright Holder Associated with a Document and Denial

N=3,770	Association	Government	Non-profit	Education	Commercial
Pearson X^2	46.45	9.98	41.94	9.26	347.91
Df	1	1	1	1	1
P	<0.001	<0.003	<0.001	<0.003	<0.001

Table 6: Denial Rate for Items with One or More Association Copyright Holders

	Accepted	Denied	Total
Items with 0 Association Rights Holders	2,955 (91.0%)	292 (9.0%)	3,247
Items with 1 or More Association Rights Holders	521 (99.6%)	2 (.4%)	523
Total	3,476 (92.2%)	294 (7.8%)	3,770

Table 7: Denial Rate for Items with One or More Commercial Copyright Holders

	Accepted	Denied	Total
Items with 0 Commercial Rights Holders	2,870 (96.4%)	107 (3.6%)	2,977
Items with 1 or More Commercial Rights Holders	606 (76.4%)	187 (23.6%)	793
Total	3,476 (92.2%)	294 (7.8%)	3,770

Table 8: Denial Rate for Items with One or More Education Copyright Holders

	Accepted	Denied	Total
Items with 0 Education Rights Holders	3,251 (91.9%)	288 (8.1%)	3,539
Items with 1 or More Education Rights Holders	225 (97.4%)	6 (2.6%)	231
Total	3,476 (92.2%)	294 (7.8%)	3,770

Table 9: Denial Rate for Items with One or More Government Copyright Holders

	Accepted	Denied	Total
Items with 0 Government Rights Holders	3,174 (91.8%)	284 (8.2%)	3,458
Items with 1 or More Government Rights Holders	302 (96.8%)	10 (3.2%)	312
Total	3,476 (92.2%)	294 (7.8%)	3,770

Table 10: Denial Rate for Items with One or More Non-profit Copyright Holders

	Accepted	Denied	Total
Items with 0 Non-profit Rights Holders	2,476 (90.5%)	261 (9.5%)	2,737
Items with 1 or More Non-profit Rights Holders	1,000 (96.8%)	33 (3.2%)	1,033
Total	3,476 (92.2%)	294 (7.8%)	3,770

Based on the indications that copyright holder type, publication status, and date category of an item were significantly correlated with an item being denied, a logistic regression model was tested integrating the different predictors. Overall X^2 for the model suggested that some of the predictors in the model were important ($X^2(7) = 330.53$, $p<0.001$). Predictors that exhibited a relatively low correlation with denial in an early iteration of the model (item has at least one non-profit, government, or education copyright holder) were removed from the final model. The approximate R^2 (Nagelkerke) was 0.22, suggesting that roughly 22% of the variance in denial was explained by these predictors. The Hosmer and Lemeshow test suggested a good fit of the model ($p>0.83$). Controlling for the other predictors, if an item was published, the odds of denial were 63% lower than for unpublished items. Even more striking, holding all other predictors fixed, when an item had at least one commercial copyright holder, the odds of denial increased by 9000% compared to items without a commercial copyright holder. Table 11 shows the odds ratios and confidence intervals for each predictor in the model.

Table 11: Odds Ratios and Confidence Intervals for Model Predictors

		Exp(B)	95% Confidence Interval for Exp(B)	
			Lower	Upper
Step 1[a]	1991-1995			
	1941-1985	.890	.287	2.761
	1986-1990	.780	.495	1.230
	1996-2000	.534	.390	.730
	2001-2005	.858	.512	1.437
	Published	.368	.274	.494
	has at least 1 Association copyright holder	.086	.021	.349
	has at least 1 Commercial copyright holder	9.151	6.833	12.255
	Constant	.078		

Using this model, one can calculate the odds that a particular item will be denied. For example, the probability of denial for an unpublished item created in 1989 by a commercial copyright holder would be calculated as follows:

$$P(denial) = \frac{\exp^{(-2.549-.117*0-.248*1-.628*0-.154*0-1*0-2.457*0+2.214*1)}}{1+\exp^{(-2.549-.117*0-.248*1-.628*0-.154*0-1*0-2.457*0+2.214*1)}}$$

=.35582, or the item has a roughly 35% chance of being denied permission for display (when only taking into account the factors in the model).

It bears repeating that the vast majority of responses to permissions requests were of the type “grant permission for all” or “deny permission for all,” with very few responses at the level of individual items (e.g. “grant permission for some”). This might indicate that the particular details of items were less important than the types of copyright holders associated with them, a premise strengthened by the very strong correlation between copyright holder type and the denial status of items.

Discussion

Based on the results of the current study, which copyright-related factors should archives consider when evaluating whether or not to digitize a particular collection for online, open access? As shown by this and other projects, cost is a major factor in mass digitization efforts, and the Cohen data demonstrate that tasks related to copyright permissions account for a significant portion of staff time. Eighty-five percent of staff time was dedicated to copyright permissions activities. While item-level description of materials was a significant contributor, far more significant was the time required to locate, contact, re-contact, and negotiate with rights holders. In fact, Cohen Project staff spent on average one hour and ten minutes per copyright holder in their attempt to obtain permission from rights holders.

The Cohen Project benefitted greatly from a dedicated source of funding. This funding supported multiple staff members who worked 20-30 hours a week on the project, a good portion of which was spent following up with rights holders. And yet, even with these resources, staff faced the same obstacle noted by Covey and George: a large number of requests (32%) were met with non-response. This fact, in conjunction with data that show that if there is no response after approximately four months, there is likely not to be one at all, might encourage archivists to place limits on the length of time they invest in the process.

The high rate of non-response, coupled with University of Michigan Library's policy that allowed the posting of only those items that were granted explicit permission, resulted in the exclusion of over 1,500 items from the online collection. Just over 30% of copyrighted items in the collection are blocked from researchers despite the huge investment of time and effort put into requesting permission and evidence that suggests that most responses to permissions requests (91% in this study) are to grant permission to display archival materials.

For archives considering mass digitization in their repositories, the data in the current study indicate several important factors to keep in mind. First of all, as indicated in the Cave, Deegan, & Heinink (2000) report, collections with a higher document to copyright holder ratio

will probably cost less to usher through the rights process than collections with a low document to copyright holder ratio. Because so much time is required to contact and negotiate with rights holders, spending that time on rights holders with many documents in a collection yields more than spending time seeking permission from rights holders with just one or two items in a collection.

Second, Cohen data demonstrate that, while very few rights holders deny permission, commercial copyright holders are more likely to deny permission than others. This means that repositories might only be able to display a small portion of collections with many corporate third party rights holders. In addition, corporate and government copyright holders are the least likely to respond to permissions requests. For repositories that equate non-response with denial, this will also have an effect on the portion of a collection available for online display.

Lastly, and perhaps more importantly, archives must reconsider the appropriateness of a policy that equates non-response with denial. This approach, commonplace in the archival profession as indicated by Dryden's study (2008), reflects an unwillingness to accept any level of risk with regard to copyright litigation. There is some risk associated in putting documents online without permission, even when the repository has exercised what it has determined to be due diligence. But, as these data have shown, there is a certain cost to restricting oneself to only posting materials with a definitive "granted permission" response, even after a significant amount of effort has gone into requesting permission from the rights holder. In the Cohen Collection, 18% of the copyright items could not be displayed due to non-response from rights holders, and another 12% could not be displayed because staff could not identify or locate the rights holders. Had the Library taken a less risk adverse approach, investing as much time in contacting rights holders, but informing them in request letters that non-response would result in posting the digitized items online until informed otherwise, 1,981 more items might be in the online Cohen Collection now. Knowing upfront that a significant number of requests will be met with non-response, and that few rights holders deny permission when they do respond, is it wise

to invest so much in the process, only to treat the associated items as good as denied? Once a repository has made the decision to invest in the process, it should consider whether a more reasonable approach to non-responses and the inability to identify and locate rights holders might be to post those items until a rights holder requests it take them down.

Future Research

This type of research would benefit from similar studies with different types of archival collections. There may have been qualities of a collection of AIDS research papers and the associated types of people and organizations represented by it that made it somehow special in terms of the copyright permissions process. Additionally, because the model for predicting the probability of a copyrighted item being denied permission for online display accounts for roughly 22% of the variance in response, there are probably additional variables that could be added to make the model stronger. For instance, document type or genre may have a significant effect on denial status (e.g. are emails more likely to be denied than meeting minutes?). I initially planned to analyze such an effect, but the fifty plus Getty Art & Architecture Thesaurus-based categories employed in the Jon Cohen Project were too unevenly and often too sparsely populated to yield any meaningful correlations in the data gathered. There might be more helpful, higher-level genre categories that could be applied to future collections to aid in this type of analysis.

Finally, the Cohen Project was guided by a policy that only allowed online posting of documents that had been explicitly approved by copyright holders. This may have had an effect on the response rate to requests. It would be interesting to compare the Cohen response rate to that of a project that employed a different approach. For instance, if permissions letters informed rights holders that non-response would result in their items being posted online until informed otherwise, would more rights holders respond? Would the rates of accept and denial change significantly?

Acknowledgements

I would like to thank Professor Margaret Hedstrom and Professor Elizabeth Yakel for their extensive and generous help with this research and in reviewing the manuscript. I am also indebted to members of the Archives Research Group at University of Michigan for their gracious and helpful feedback; the Center for Statistical Consultation and Research (CSCAR) at the University of Michigan, especially Brady West, for assistance with the statistical analyses; Alissa Centivany for her help understanding some of the intricacies of copyright law; Jon Cohen; University of Michigan Library; and the Cohen Project staff. This research was funded by a grant from the John D. Evans Foundation.

References

Besek J. M. (2003). *Copyright issues relevant to the creation of a digital archive: A preliminary assessment*: National Digital Information Infrastructure and Preservation Program, Library of Congress.

Cave M., Deegan, M., & Heinink, L. (2000). Copyright clearance in the refugee studies centre digital library project. *RLG DigiNews, 4*(5).

CENDI Copyright Working Group (2008). Frequently Asked Questions about Copyright. Retrieved April 11, 2010 from http://www.cendi.gov/publications/04-8copyright.html.

Covey D. (2005). *Copyright and the universal digital library*. International Conference on the Universal Digital Library. Hangzhou, China. Oct. 2005.

Covey D. T. (2005). *Acquiring copyright permission to digitize and provide open access to books*: Digital Library Federation Council on Library and Information Resources.

Dryden J. (2008). Copyright in the real world: Making archival material available on the internet. *College & Research Libraries News, 69* (7).

George C. A. (2002). *Exploring the feasibility of seeking copyright permissions*: Carnegie Mellon University Libraries.

George C. A. (2005). Testing the barriers to digital libraries. *New Library World, 106*(7/8).

Hirtle P. B. (2001). Unpublished materials, new technologies, and copyright: Facilitating scholarly use. *Journal of the Copyright Society, 49*, 259-275.

Hirtle P. B. (2003). Digital preservation and copyright. Retrieved November 1, 2008, from http://fairuse.stanford.edu/commentary_and_analysis/2003_11_hirtle.html.

Hirtle P. B. (2006). Digital access to archival works: Could 108(b) be the solution? Retrieved November 1, 2008, from http://fairuse.stanford.edu/commentary_and_analysis/2006_08_hirtle.html.

Hirtle, P. B. (2008). Copyright renewal, copyright restoration, and the difficulty of determining copyright status. *D-Lib Magazine, 14*(7/8).

Hughes L. (2004). *Digitizing collections: Strategic issues for the information manager*. London: Facet Publishing.

Lee S. (2001). *Digital imaging: A practical handbook*. New York: Neal-Schumann Publishers.

Pritcher L. (2002). Ad*access: Seeking copyright permissions for a digital age. *D-Lib Magazine, 6*(2).

RLG (1998). *Worksheet for estimating digital reformatting costs*.

Sitts M. (2000). *Handbook for digital projects: A management tool for preservation and access*: Northeast Document Conservation Center.

Stanford University Libraries (2007). *Copyright Basics FAQ*. Retrieved October 2, 2009 from http://fairuse.stanford.edu/Copyright_and_Fair_Use_Overview/chapter0/0-a.html#1.

United States Copyright Office (2006). How to investigate the copyright status of a work Retrieved November 1, 2008.

United States Copyright Office. Defintions FAQ. Retrieved October 2, 2009 from http://www.copyright.gov/help/faq/faq-definitions.html.