The Copyright Review Management System, Its Development and Process

University of Michigan Copyright Office
Copyright Review Management System

Judith Ahronheim
Head, Electronic Resource Access Unit
Library Technical Services
University of Michigan
jaheim@umich.edu
ORCID: 0000-0002-0259-280X

Kristina Eden
Associate Librarian
Michigan Publishing, Copyright Office
University of Michigan
keden@umich.edu
ORCID: 0000-0003-4794-3906

10/1/2013
THE COPYRIGHT REVIEW MANAGEMENT SYSTEM, ITS DEVELOPMENT AND PROCESS
Judith Ahronheim, Kristina Eden

ABSTRACT
Since its inception in 2008, HathiTrust has grown to be one of the largest collections of digital documents in the world. Identifying which of its multitude of documents may be made freely available for users has become a collaborative project of great scope. Because such a large population of materials is involved, our experience of developing policies and procedures for identifying public domain materials should have application for a wide variety of other projects.

INTRODUCTION
The digital document collection known as HathiTrust was born of agreements among several academic libraries and Google in 2008 and was preceded by agreements the University of Michigan and others made with Google in 2004. At that time academic libraries with digitization projects were looking at a very long timeline for projects and were generally selective about which materials to digitize. Most investigation of rights occurred prior to digitization and took significant time and effort.

The association with Google produced a sea change in the volume of materials that could be digitized and in the ambitions of academic libraries to share their collections more widely online. As the fruit of the Google partnerships appeared on servers, academic caretakers came to the realization that “it was possible to scan millions of pages in a ridiculously short period of time, making the possibility of digitizing large collections a reality.” With the production of this huge collection of digitized content, an opportunity became apparent to make a large body of public domain works sharable. Was there a way to identify materials in the public domain that could be applied to a large population at a rate that approximated that of the scanning?

In 2007 the University of Michigan Library began developing the foundation for what would become the Copyright Review Management System (CRMS). A grant proposal was submitted to the Institute for Museum and Library Services that set a goal of 60,000 public domain works identified within a three year period. This IMLS National Leadership grant was awarded in 2008 and completed in November 2012 with over 170,000 books reviewed and 87,000 determined to be in the public domain. A second National Leadership grant was awarded for 2012-2014 to continue the work of CRMS and additionally look at works published in the United Kingdom, Canada, and Australia.

The initial effort to identify public domain items focused on materials published in the United States prior to 1923 because it was possible to use information found in the bibliographic record to make an automated determination regarding copyright status. This was a fairly well established means used in a number of earlier digitization projects, albeit rarely applied on the HathiTrust scale. While dependence on data in the bibliographic record made a not entirely reliable specification, since bibliographic aims differ somewhat from copyright aims, it was a method that allowed for machine processing. Details on the specification used may be found on the HathiTrust site. The bibliographic record was also used as a rough means of identifying U.S. federal government documents and sequestering them for additional filtering and human review before they were declared to be public domain documents. This process occurred outside the CRMS project.
<table>
<thead>
<tr>
<th>Publication Date</th>
<th>Other Condition</th>
<th>Copyright Status in Rights Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>1922 or Earlier</td>
<td></td>
<td>Public Domain</td>
</tr>
<tr>
<td>1923 or Later</td>
<td>Not U.S. Govt. Publication</td>
<td>In Copyright</td>
</tr>
<tr>
<td>1923 or Later</td>
<td>U.S. Govt. Publication</td>
<td>Public Domain</td>
</tr>
<tr>
<td>No Date Information</td>
<td></td>
<td>Undetermined</td>
</tr>
</tbody>
</table>

U.S. works published from 1923 through 1963 had to comply with formalities in order to achieve copyright protection. The formalities included publication with proper copyright notice and renewal of the copyright registration with the U.S. Copyright Office after 28 years. The Library Technical Services Division at the University of Michigan Library began “manual” copyright review work in spring of 2007 as a pilot project with the intention of increasing access to works in the public domain. There were approximately 5 staff members contributing time that amounted to 1.5 Full Time Equivalent (FTE). The primary focus was on books, not periodicals; and considerable effort went into developing guidelines and detailed use cases.9

**Improving the Process and Developing an Interface**

There were three issues with the original “manual” copyright review process: organization, access, and reliability. Organization and tracking of data was first managed with numerous spreadsheets which proved unwieldy as the project increased in size. Access to information about prior decisions was difficult to retrieve easily, and the process led to questions about the reliability and consistency of each reviewer doing work in isolation. Checking for copyright renewal had to be done using individual scanned volumes of the U.S. Copyright Office’s Catalog of Copyright Entries (2 per year, requiring a search of six volumes to determine if a renewal existed) that had not been converted to searchable text. This proved to be excessively time consuming. The appearance of the Stanford Copyright Renewal Database, also in 2007, presented a more efficient method of checking for renewals.10 The existence of the Stanford Copyright Renewal database is one of the primary reasons why the CRMS project is feasible. It enables an accurate search of renewal records online in a matter of minutes. Conversely, a sample set of volumes sent to the U.S. Copyright Office as part of the grant’s audit process required months for staff there to search the physical copyright renewal records.11

Receipt of the IMLS grant funded an effort that was able to create the Copyright Review Management System interface, a more efficient system for managing copyright determinations. This helped prioritize workflow, incorporated useful tools, instituted a double-review standard to ensure reliability of results, and made it possible for other institutions to join the work.12

Version 1.0 of the CRMS interface was released in July 2009 and “manual” work ceased. As part of the new CRMS process, rights determinations exported to HathiTrust went through at least two independent copyright status reviews. In the summer of 2010, project staff trained additional reviewers at Indiana University, University of Minnesota, and University of Wisconsin, increasing the number of reviewers to 18 people across four institutions. The productivity began to increase dramatically in the fall of 2010.
Development of the CRMS system was an iterative process with changes being worked back into the database, workflow, and web interface as project members gained experience and analyzed the results. Since there was a promised grant outcome to identify 60,000 public domain books, increasing the tool’s efficiency to get through a large number of volumes during the grant period was a high priority.

**HOW THE PROCESS WORKS**

Each volume is reviewed by two independent reviewers. If the two agree, then the rights determination is a match. If they disagree, then it is a ‘conflict’ and an expert reviewer must arbitrate the decision. Final determinations are recommended daily to the Rights Database which interacts with HathiTrust to produce data suggesting which books are public domain and therefore should be made fully viewable online. Data indicates that the double review process resulted in less permissive (more conservative) determinations, which was a top priority for this project. From 2008-2012, only about 69% of determinations were made without conflicts. Adding in the double review caused the review rate to slow, and greatly increased the amount of work put upon expert reviewers. Because of the 60,000 book goal, this had to be addressed very quickly.

The solution was to standardize the order of decisions made by reviewers. Having standardized steps in the determination workflow produces a more consistent assignment of rights/reason codes, while not making significant changes in the ultimate public domain decision. Since all reviewers are taught to follow the same sequence of steps in making decisions, works that have a mix of factors affecting their copyright status are more likely to receive similar coding. This is a solution to the large number of slightly differing reviews (“conflicts”) that require expert resolution even though the end result is the same.
FIGURE 2
The CRMS Data Flow
EFFICIENCY IN A LARGE-VOLUME WORKFLOW

CRMS has been designed to manage a large volume workflow, ensure an accurate and complete decision trail and limit to securely authorized personnel access to potentially copyright protected book scans. Coding underlies the system, controlling the workflow and the priority in which records are queued up. A status number is assigned to each volume as it makes its way through the process. This status number is used to determine who is able to see the volume. In a standard route through the process, a volume begins at status 0 and after being reviewed by two reviewers it receives either a 2, 3, or 4 status according to decisions made. This either routes it into a conflict queue, marks it as a provisional match, or assigns a final determination status. If an expert takes part in the process, then it receives status 5. This simplifies our ability to manage thousands of records and look at historical decisions in light of any number of criteria.

The automated routing also ensures proper procedure is followed without needing to rely on human memory. Once a volume has been looked at by one reviewer, it will be sent to another automatically by the system. Records are served randomly, keeping the process objective rather than allowing reviewers to select the book they want to work on. Also, by virtue of the assigned status code, reviewers are locked out of old volumes and unable to go back in and retroactively change rights codes in the database. A grace period is allotted until the end of the day for reviewers to change their mind about a decision, pull up a record, and make a correction. Overnight the records are matched up, removed from the reviewer’s “unprocessed records” queue, and moved on to a next stage.

Having a structured framework to the interface provides prompts and cues in real time. If reviewers attempt to enter an illogical decision, there are pop-up scripts to question their action or require them to enter an explanatory note. This provides certainty that everyone is seeing the same thing and using the same resources. The interface is designed to reinforce the foundational methodology, and decision options are standardized as much as possible to ensure staying within scope and the legal framework.
All the tools needed to make a review are directly accessible within the review interface. Putting all the resources into a single interface brings everything to the fingertips of the researcher. It eliminates time wasted in managing several resources in other browser windows. It also ensures that reviewers working remotely from multiple institutions are all seeing the same information.

A basic bibliographic stub containing volume ID, title, author, and publication date is pulled into the interface, and a toggled view to the MIRLYN catalog record (the University of Michigan’s OPAC and the current HathiTrust database of record) is provided. In addition, a wiki documenting policy decisions made in the course of review and copies of all training documentation are available directly within the interface.

Before reviewers even see an incoming volume from HathiTrust, filters have been applied to ensure that it falls within scope. From the HathiTrust pool of books, volumes that receive a category ic/bib (in copyright by virtue of bibliographic data) are then further filtered by format = book, place of publication = U.S., and date of publication = 1923-1963, for example. Additional filters remove non-English languages, translations, and dissertations. These filters enable the project to pre-select a pool of works most likely to result in a fruitful review and minimize the number of volumes that would be out of scope. The filters are continually being modified as reviewers provide feedback to the developer. The filters, however, are driven by catalog data and so the automatic filtering is only as accurate as the dates and fixed fields in the catalog record.
CODING TO MANAGE RIGHTS INFORMATION

The goal of the CRMS process is to assign a right/reason code to each evaluated work, and HathiTrust uses this code to help regulate user access to the work. Over time the set of codes has evolved to become a means for storing a simplified, machine-actionable set of information about the copyright status of a work.

A reviewer begins by examining the digitized scan of the work for obvious factors that would eliminate it from consideration: foreign publication, translation of foreign work, etc. Then he launches a search of the Stanford database by either author or title. Data for the search is automatically drawn from the catalog record. If a renewal record is not found in the first try, the reviewer also must search on variations in the name and title in order to try for a comprehensive verification that no record of renewal exists in the Stanford database. A panoply of search techniques such as truncation, removal of punctuation, and search term granularity is taught during training to aid reviewers in finding all relevant records. The objective is to satisfy every possibility that might locate a renewal record. A known bug in the database is that truncating an author’s name that begins with A or R will result in an error message and in this case, reviewers work around this issue.

If a copyright renewal is located and there are no other complicating factors in the book, then the reviewer will paste the renewal record number from Stanford into a field in the CRMS interface. A CRMS script then automatically runs to retrieve the renewal date field from the Stanford record. The volume is then assigned the code ic/ren (in copyright/renewed).

CODES ASSIGNED BY CRMS-US (UNITED STATES)

PD is the rights code representing public domain works. The second part of each code represents the reason for finding the volume to be in the public domain.

**pd/ncn: Public domain/No copyright notice**

One of the criteria for copyright renewal of U.S. publications between 1923-1963 was that a copyright notice be printed on the book. From the start of the CRMS project in 2009, we noted which of the books lacked a copyright notice. Through the course of investigation, it was noticed that the digitized scans were sometimes missing front matter pages where a copyright notice might have existed. As it is difficult to prove a negative, the absence of a copyright notice on the book scan became a less desirable method for checking copyright status. Although there may be a fair number of books that could reasonably be held in the public domain for failure to comply with formalities, this condition is less assured than searching for a record of copyright renewal and as of July 10, 2013 the pd/ncn reason code has been discontinued as an option in the interface and is no longer being used.

**pd/ren: Public domain/Renewal checked**

Books which make it past the initial checks are searched in the Stanford database for a copyright renewal registration. This code is assigned if no renewal record is found.

**pd/cdpp: Public domain/ Publication date prior to 1872**

**pdus/cdpp: Public domain U.S. only/ Publication date prior to 1923**

In some cases the copyright date of a book is found to be pre-1923. These books are marked with a code to indicate the public domain status due to copyright date rather than any criteria related to the 1923-1963 publication requirements. This reason code is also applied to books which are exact reprints of a pre-1923
publication. Reprints of a public domain item will also be in the public domain unless there is newly added copyrightable content.

IC is the rights code representing in-copyright works. The second part of each code represents the reason for finding the volume to be in-copyright. Books assigned this code are confirmed to be in-copyright because a renewed copyright registration was found.

ic/cdpp: In copyright/ Publication date post 1963
ic/crms: In copyright/ Multiple reasons
A minor amount of books were found to be published post-1963, a period which is outside the scope of the project. The CRMS filters normally catch these based on bibliographic information, however in some cases such as set records for monographic series, publication dates do not fully reflect the data for individual volumes. This is why it is important for the reviewer to visually inspect the title page and verso to confirm the publication date, since books published after 1963 are subject to different copyright terms.

UND is the rights code representing “undetermined” works. The second part of each code represents the reason for leaving the copyright determination undetermined.

und/nfi: Undetermined/ Needs further investigation
und/crms: Undetermined/ Multiple reasons
“Undetermined” is a catch-all reason code use to identify the presence of complexities causing the reviewer to put a book aside for later review. Rather than engage in a definitive process for every book passing through the queue, it was deemed more efficient, given the total body of works, to first focus on readily identifiable public domain materials and set aside complex cases until later. The data collected on these “undetermined” books includes a note category selected by reviewers and typed notes indicating the nature of the problem found. In some cases the two reviewers agreed that a book should be “undetermined” but had differing reasons for doing so. The amount of time needed to research this category of books could possibly increase exponentially which is why they are set aside for later review, thus the separate “crms” reason code.

Once a volume has completed the review process, the rights/reason code is exported to the rights database. This is the point at which a volume would be opened up in HathiTrust if the rights/reason code indicates that the volume is in the public domain. Volumes still in-copyright or needing further investigation merely have their rights/reason code updated indicating some copyright research has taken place beyond the automated “bib” determination, but the volume remains closed within HathiTrust.

THE CRMS INTERFACE AS A TOOL FOR EFFICIENT COLLABORATION
Working within the CRMS interface allows asynchronous work in whatever segments of time reviewers have available to fit their schedule. Reviewers are contributing their cost-share time in between their other duties at university libraries. Many are librarians and catalogers, and part of the nature of their work is being interrupted or finding short blocks of time between meetings and other obligations to do some review work. If they need to get to a reference desk shift or answer a phone call, at any time during the process a reviewer can cancel out of a request. The volume is then placed back into the queue to be served up to whichever next reviewer has time available. Reviewers are also able to place a hold for up to three days on a record, giving them time to ask and receive an answer to a question they might have. This efficient routing and
management of records allows reviewers to work whenever they want without being dependent upon the work of anyone else or worrying about the competing factors for their time.

Within the CRMS interface there is no limit on how many people can be working simultaneously, unlike working with shared spreadsheets where only one person can enter data at a time. All of the reviewers can work at a time that is convenient to their schedule without needing to coordinate their time with others. Opening up a record for review will systematically lock everyone else out of it for as long as one person has it open. This prevents duplicative work if many people are in the system at the same time. It also eliminates the problem of file versioning and keeping track of spreadsheets of data.

CRMS AS A TOOL FOR DOCUMENTATION AND DUE DILIGENCE
The CRMS ensures that a perpetual history of reviewer decisions is retained within the interface, ameliorating the issue of lost data which could happen when someone transfers to a new job or leaves a department. The system is intended to produce not only copyright determinations, but responsibly document how every single determination was reached and allow it to be retrieved and interpreted even after the project has been completed.

Each decision made by a reviewer regarding a particular volume is stored as a separate data entry referred to as a ‘determination’. Historical data is searchable by reviewer name, date of review, free text search of the notes field, and other items. This is used both by supervisors and reviewers themselves to look at work history, see all determinations given for a particular volume and observe how an expert adjudicated any conflicting determinations. Because there is a record of determinations in one constantly updating database, a project supervisor is able to search the data to answer questions and provide feedback on reviewers’ performance.

COSTS AND STAFFING ISSUES
Personal statistics are displayed to reviewers within the CRMS interface. On the following day, reviewers can see whether their reviews achieved a match or became a conflict, and how their statistical averages compare to the others on the project. Viewing personal statistics is often a motivator for reviewers and helps them monitor time commitments to the cost-share on the grant, as well as providing feedback on the accuracy of their work.

Throughout the project certain costs were assumed for start-up, training and ongoing development which have proven accurate. The costs are not insubstantial and for anyone intending to design and carry out a copyright research project, our experience can be valuable.

Depending on the scale of the project, daily management might be carried out by one person or several, but we found that skill sets are best spread over several individuals to encourage redundancy, share the workload, and provide stability to the project in case of job change. The leadership team brought together personal expertise in instructional design for training, operational management and report writing, and step-by-step documentation. It is prudent to hire a dedicated developer, because even in the project maintenance phase, daily issues occur that require immediate attention to avoid loss of productivity. There must be consistent support for a homegrown interface, particularly one utilizing access to diverse external systems, and responsiveness to upgrades and debugging. And with a project gaining access to scans of digitized books such as HathiTrust, it would be unthinkable to operate without a security expert guarding the gateway. An
extremely rigorous level of security on this project ensures that only properly authenticated reviewers working from their assigned computer terminal can see scans of the books. In fact, the security is so tight that newly authorized reviewers often have difficulty being added to the system.

To set the foundation for all decisions, there must be strong legal research. The formation and oversight of legal assumptions setting the basis of a copyright project should not only be researched and approved by legal counsel, there should also be an advisory board of copyright experts to inspect and critique the assumptions prior to enactment. Since copyright is so complex, relying on this body of advisors to validate the process is additional assurance that nothing has been overlooked. Without this oversight, it is possible to be too conservative or too open.

The decision process for reviewers turned out to be more complex than was originally scoped. Beyond simply looking for a renewal record, a practiced eye is helpful to identify periodicals, non-Class A works, and authors of non-U.S. origin. The corollary of underestimating the complexity is that the amount of expert time needed to resolve conflicts was also underestimated in the beginning. Originally the project manager was intended to be the sole arbitrator, but it became apparent that additional experts would need to be trained to reasonably manage the workload.

Finally, if the work is a collaboration among geographically distant universities, it is important to budget for travel expenses related to training. At least two days of in-person training was the way chosen to kick-start the project and maximize consistent work product. Follow-up booster sessions or addition of new trainees can be done by remote operations, but there is nothing to substitute for an extended period of face-to-face discussion, questioning, and a supervised trial run for everyone. Likewise, although personal schedules of high level officers can be difficult to coordinate, hosting a face-to-face meeting for the advisory board is a catalyst for productive copyright discussion and picking apart base assumptions. It is wise to budget travel and hosting expenses for bringing the copyright advisory board together for a working meeting.

**CONCLUSION**

The CRMS interface has proven to be an efficient and productive tool for examining a large number of scanned items for public domain status in a highly collaborative and distributed environment. In a short period of time it has made a 142,614-volume dent in the black box of copyright ambiguity that forms the 1923-1963 period in U.S. publication. Its success was an inspiration and model for CRMS-World, a new interface which is now being used to examine works published in United Kingdom, Canada and Australia. As additional participants come forward, it can also be used to examine works published in other parts of the world as well. It is hoped work such as that done through CRMS will propagate and continue to open access to public domain materials for people all over the world.

---
