January 2003

Copyright Permission: Turning to Dust or Digital

Denise Troll Covey
Carnegie Mellon University, troll@andrew.cmu.edu

Published In
International Journal of the Book, 149-158.

Copyright Permission: Turning to Dust or Digital

Denise Troll Covey
Associate University Librarian, Carnegie Mellon
Future of the Book Conference – Cairns, Australia – April 2003

Understanding the Issues

Our cultural and intellectual heritage is in danger. Millions of books printed on non-acid-free paper are turning to dust on library shelves. Copyright seriously impedes their preservation through digitization because permission rights must be negotiated for each title. Even if the books are digitized, current practice is to restrict their use through licensing provisions or digital rights management technologies that can trump otherwise legal uses of the materials. Lack of commitment to perpetual access threatens digital books with disappearance or darkness in an inaccessible archive. The private interests of copyright holders appear to override the public good of free and open access to information.

The core values of librarianship are endangered by this state of affairs. Our cries for equitable access, stewardship, service, and privacy (as handmaidens of intellectual freedom, democracy, and literacy) seem to fall on deaf ears. Government legislation suggests collusion with commercial industries, and the once hallowed public domain seems to wither from starvation. The United States government passed the Copyright Term Extension Act (CTEA) in 1998. Otherwise known as the Sony Bono or Mickey Mouse Act, because the primary lobbyists were Hollywood media conglomerates and the Walt Disney Corporation, this law extended copyright to the life of the author plus 70 years, or 95 years for a work for hire. Attempts to have the CTEA declared unconstitutional failed in 2002 when the U.S. Supreme Court reviewed the case and ruled that the law did restrict the copyright monopoly to a "limited time."

Approximately 95% of the books ever published are still in copyright, but fewer than 3% of the published books are still in print. This means that 92% of the world's books are neither generating revenue for the copyright holder nor easily accessible to potential readers. This is absurd. If copyright cannot be shortened, the only solution to the current situation appears to be to negotiate permission to digitize and provide open access to copyrighted works.

Apart from concerns about preservation, librarianship, and the copyright absurdity, lack of a sufficient body of quality materials readily available online is creating significant problems in higher education. Students and faculty look online first when they need information because of the speed and convenience of Internet access. Recent research conducted by Outsell Inc. and the Online Computer Library Center (OCLC) indicates that faculty and graduate students typically turn to library-provided electronic resources to do their research.[1] These materials reside in the

[1] The Outsell study is reported by Amy Friedlander, *Dimensions and Use of the Scholarly Information Environment: Introduction to a Data Set* (Washington DC: Council on Library and Information Resources publication 100, October 2002). Available: <www.clir.org/pubs/reports/pub110/contents.html> See also *How Academic Librarians Can Influence Students' Web-Based Information Choices* (Dublin, Ohio: OCLC White Paper

"invisible" or "deep web," which is said to be 500 times larger, growing much faster, and far superior (in quality of content) to its counterpart, the "visible" or "surface web."[2] Most undergraduate students (72%), however, turn to popular Internet search engines like Google or AltaVista to do their assignments. These popular search engines index only the content available on the surface web. Faculty are concerned about the quality of the resources that undergraduate students are using to complete their assignments because only 6% of surface web content is appropriate for student academic work, because no single Internet search engine indexes more than 16% of the surface web,[3] and because the trend is for the results retrieved by popular search engines to be ranked by fees paid by advertisers or sponsors rather than by relevance to the user's query.[4] Estimates suggest that an Internet search conducted using a popular Internet search engine actually searches only 0.03% (1/3000) of the content available on the entire (surface + deep) web.[5] The bottom line is that lack of quality resources on the surface web is having a negative impact on the quality of student learning.

Meanwhile, the increasing availability and use of online bibliographic databases, the increasing number of scholarly publications, and the increasing cost of library materials have created a situation wherein libraries are spending more money but purchasing fewer materials. Interlibrary loan is increasing, with many libraries underwriting the cost. However the turn-around time is often inadequate for both the rapid-paced, highly competitive research conducted by faculty and graduate students and the shorter deadlines and last-minute efforts of undergraduate students. Consequently, user satisfaction is decreasing. Lack of speedy access to quality resources is having a negative impact on the timeliness and success of academic work.

Beyond the boundaries of these problems, tremendous disparity exists across the nation and around the world in the size and accessibility of library collections. Some single institutions, like Harvard and Yale, have more books in their libraries than some entire states have in all of their libraries combined. In our rapidly changing world, lifelong learning and access to books have become essential to employment, health, peace, and prosperity. The democratization of knowledge and empowerment of a global citizenry require equitable access to information.

From the perspective of stewarding our cultural and intellectual heritage, achieving the mission of higher education, and engendering democracy, something is seriously awry with the current

on the Information Habits of College Students, June 2002). Available: <http://www2.oclc.org/oclc/pdf/printondemand/informationhabits.pdf>

[2] Marcia Mardis. "Uncovering the Hidden Web, Part I: Finding What the Search Engines Don't." *ERIC Digest* EDO-IR-2001-02 (October 2001). Available: <http://www.ericit.org/digests/EDO-IR-2001-02.shtml>

[3] Steve Lawrence and Lee Giles. "Accessibility and Distribution of Information on the Web." *Nature* 400 (1999): 107-109. Summary of findings available at <http://www.wwwmetrics.com>

[4] With its recent acquisition by Overture, AltaVista joined the growing number of search engines operating under an advertiser or sponsor-driven business model for ranking search results. Barbara Quint. "Overture Acquires Two Major Web Search Engines." *Information Today, Inc. Online* (March 3, 2003). Available: <http://www.infotoday.com/newsbreaks/nb030303-1.shtml>

[5] searchWebServices.com Definitions. See "deep Web," http://searchwebservices.techtarget.com/sDefinition/0,,sid26_gci558034,00.html dated June 5, 2001.

situation. Steps must be taken to enhance publisher understanding of the value of providing open access to copyrighted works.

Exploring the Feasibility

In 2001, Carnegie Mellon University Libraries completed a feasibility study conducted to determine the likelihood of publishers granting permission to digitize copyrighted books and offer them free-to-read on the surface web. The study, based on a statistically valid random sample of books in Carnegie Mellon's library catalog, revealed that locating copyright holders is time-consuming, expensive, and often unsuccessful. Fewer than half of the publishers located in the study responded to letters of inquiry. Among those who did respond, the overall success rate for getting permission to digitize a book was 22%, though the success rate varied with different types of publishers, ranging from 45% for scholarly associations to 12% for commercial publishers. But even when permission was granted, seldom was it given to offer the digitized book free-to-read on the web. More often than not restrictions or fees were applied. For example, access was to be restricted to the Carnegie Mellon community, a permission fee was levied of $100 to $300 per book, or permission was granted for only a couple years, after which the book would have to be removed from the web. The feasibility study revealed that it is indeed possible to secure permission to digitize books and offer them free-to-read on the web, but that more effective strategies for persuading copyright holders and containing costs would be required to pursue copyright permission for open access on a larger scale.

Introducing the Million Book Project

Despite the low success rate in securing copyright permission in the feasibility study, confronted with the threats to our cultural and intellectual heritage, and concerned about the quality of education and the democratization of knowledge, Carnegie Mellon decided to forge ahead. In 2001 the School of Computer Science and the University Libraries launched an international project – more like a campaign – to digitize a million books in five years and offer them free-to-read on the surface web. The National Science Foundation (NSF) provided $3.6 million to purchase scanning equipment and support travel required for project management. The government of India is providing $1.5 million for labor. University libraries in the United States have joined the Million Book Project, as have the Online Computer Library Center (OCLC) and the Internet Archives. Several other countries have asked to join the project, including Egypt, Taiwan, and Hong Kong. Additional information about the Million Book Project can be found at http://www.library.cmu.edu/Libraries/MBP_FAQ.html.

Scanning is currently underway in India, creating well-paid jobs in an impoverished market. Scanning follows established standards for bibliographic metadata and file formats to ensure interoperability with existing systems and migration to new technologies in the future. Several scanning centers are operational and more are planned, including a "super center" in Hyderabad. A typical center has 6 to 8 Minolta scanners operating two eight-hour shifts per day. The productivity rate is 16 books per day per scanner or roughly 4000 books per year per scanner. The goal is to have 100 scanners in India, digitizing 400,000 books per year. Allowing for a generous 50% deterioration rate, the Million Book Project can be completed in five years. When a MARC record is available for a title, librarians capture the metadata from OCLC WorldCat using guest IDs provided by OCLC. When no MARC record is available, they create a Dublin Core record for the title. Scanner operators digitize the books and perform quality control post-processing to de-skew the images and remove black edges. The books will be replicated on servers around the world, indexed by popular search engines, and freely available on the surface web – accessible anywhere, any time, to anyone with an Internet connection. The University Libraries plan to register the books in the Million Book Collection in OCLC's digital registry to avoid redundant scanning of the same books. Any school, public, or academic library will be able to link their library catalog records to the books in the Million Book Collection. The Collection will support education, research, and lifelong learning worldwide.

The initial collection development meeting for the Million Book Project was held in November 2001. Librarians from fifteen academic libraries in the United States assembled with representatives of OCLC, NSF, and the Digital Library Federation (DLF). Participants agreed that the Million Book Collection would be a collection of collections, comprised of out-of-copyright books, copyright protected books, and works like government documents that cannot be copyrighted. Project partners in India agreed to digitize 200,000 indigenous Indian books, and 700,000 public domain and 100,000 copyrighted books shipped from the United States. The copyrighted books targeted for initial pursuit were the (approximately) 50,000 titles cited in *Books for College Libraries*, which is a five-volume bibliography of essential holdings for all college and university libraries. A second collection development meeting is planned for November 2003. Key project partners need to reconvene to select additional bibliographies and strategize about how to locate and acquire the books and coordinate their shipment to India. Approaching collection development for the Million Book Project through the use of existing bibliographies is an efficient, effective method for creating a collection of quality collections. Selecting books title-by-title is too expensive on a large scale, replicates work that has already been done by librarians, and is problematic for other reasons, not the least of which is contention over who decides which books are essential, based on what criteria.

Seeking Copyright Permission

Indian partners are doing the copyright permission work for their indigenous materials. The copyright permission work for books selected and shipped from the United States is centralized at Carnegie Mellon University Libraries. Applying lessons learned from the feasibility study, the University Libraries is piloting a new approach to acquiring copyright permission, an approach based on educating and providing incentives for publishers of quality academic books to grant permission to digitize their out-of-print, in-copyright books and offer them free-to-read on the web. The new approach has already secured permission from publishers to digitize thousands of books and include them in the Million Book Collection.

Letters to publishers briefly introduce the Million Book Project, explicitly state our adherence to copyright law, and describe the copyright absurdity wherein out-of-print, in-copyright books are neither generating revenue for them nor readily available to potential readers who might be willing to pay for them. We provide an overview of research indicating that

- Users want to find information online, but use it in print.[6]
- Online access increases use, including use of older materials.[7]
- Open access does not decrease revenue. It can actually increase sales. [8]

The letters ask publishers for non-exclusive permission to digitize their books and offer them free-to-read on the surface web. We explain that the delivery system will have minimal functionality, specifically that printing will not be fully supported. We offer to give publishers that grant permission preservation-quality copies of their books and the associated metadata and OCR (text file to facilitate full-text searching). To motivate their agreement, we explain that they can use these copies in added-value, fee-based services that they develop or use. They could provide a Buy button and Print-On-Demand service in conjunction with the images that would generate revenue for them from the sale of in-print and out-of-print books. We do not offer to pay – and will not pay – a permission cost per title.

Initially, our copyright approach focused on selected titles, specifically the titles cited in *Books for College Libraries* (BCL). A cursory study of the 1988 edition of BCL revealed that of the 50,000 books cited:

- 2,200 (4%) of the titles were published in or prior to 1923. These books are out of copyright and can be digitized and included in the Million Book Collection without asking permission.
- 35,500 (71%) of the titles were published between 1924 and 1964, when copyright in the United States required renewal. Evidence suggests that copyright to 80%-90% of the titles published during this period was not renewed.[9] Therefore an additional 28,000 to 32,000 of the BCL titles could be out of copyright.
- 12,300 (25%) of the titles were published after 1964. These books are copyright protected and require permission to digitize and include in the Million Book Collection.

We quickly realized that the transaction cost of checking the copyright renewal records, preparing and mailing letters listing all of a publisher's titles cited in BCL, and tracking permission per title would be prohibitively expensive. For example, Harvard University Press published approximately 1000 titles cited in BCL. Brooding over the problem, late one night I had an epiphany: rather than focusing on titles, we should focus strictly on publishers. We should treat BCL and other selected bibliographies like an approval plan for publishers. This

[6] Amy Friedlander, *Dimensions and Use of the Scholarly Information Environment: Introduction to a Data Set* (Washington DC: Council on Library and Information Resources publication 100, October 2002). Available: <www.clir.org/pubs/reports/pub110/contents.html>

[7] Kevin M. Guthrie. *Revitalizing Older Published Literature: Preliminary Lessons from the Use of JSTOR* (March 23, 2000). Available: <http://www.jstor.org/about/preliminarylessons.html>

[8] See "National Academy Press: A Case Study," by Barbara Kline Pope, *The Journal of Electronic Publishing* 4, 4 (June, 1999). Available: <http://www.press.umich.edu/jep/04-04/pope.html>

[9] In 1960, fewer than 15% of all registered copyrights were being renewed. See Barbara A. Ringer, "Study No. 31: Renewal of Copyright." In *1 Studies on Copyright*, Arthur Fisher Memorial ed., June 1960: 513-514. See additional research cited on pages 9-11 of the *Brief of Amici Curiae of the Internet Archive, Prelinger Archives, and Project Gutenberg Literary Archive Foundation*, No. 01-618 in the Supreme Court case Eric Eldred v. John D. Ashcroft.

would obviate the need to check copyright renewal records, greatly reduce the cost of preparing and mailing letters, and with each letter sent potentially secure permission to digitize and include more titles in the Million Book Collection than just those cited in the bibliographies. If only 3% of the 5600 publishers with works cited in BCL grant permission to digitize 3000 books each, the result would be 504,000 copyrighted works for the Million Book Collection. If negotiations with publishers cited in BCL achieve the 22% success rate of the feasibility study, the result could be millions of books. We would need to negotiate with India for more labor!

Our current letters ask publishers for permission to digitize and offer free-to-read on the web any of the following options:

- All of their out-of-print, in-copyright titles
- All of their titles published prior to a date of their choosing
- All of their titles published *N* or more years ago – they specify *N*
- A list of titles that they specify

Using this strategy, several publishers have granted permission and negotiations with many others are well underway. Our hypothesis has been confirmed. For example, though BCL lists only 26 titles published by the National Academy Press, the publisher granted permission to digitize all of their titles (that they hadn't already digitized) published prior to 1995. The result was permission to digitize 3,400 titles – 99% more titles than are cited in BCL. As more publishers agree to participate in the Million Book Project, the Project will no doubt already have permission to digitize many of the copyrighted books cited in the additional bibliographies that will be selected by project partners in 2003.

Results of the 2001 copyright feasibility study revealed that the success rate in seeking copyright permission varies by publisher type. Scholarly associations are almost four times as likely as commercial publishers to grant permission to digitize and offer their books free to read on the web. University presses are more than three times as likely to grant permission. Museums and art galleries are two and a half times as likely to grant permission. Using this information, copyright permission work for the Million Book Project is focusing on scholarly associations and university presses. University presses alone published approximately 27% of the books (13,528 titles) listed in BCL. Letters requesting copyright permission have already been sent to eleven of these university presses, accounting for almost half (6,224) of the books published by university presses and cited in BCL.

Tracking the Effort

Carnegie Mellon University Libraries are creating a FileMaker database to track publisher addresses, contact information, negotiations, and outcomes. Authority controls will be devised so that publisher names are consistent for reporting and assessment purposes. More recently we began developing a web-based database to capture administrative metadata for each title. Development of the administrative metadata database was precipitated by changes in how copyrighted books are acquired and shipped to India. Initial Project plans were to acquire copyright permission before shipping and digitizing any copyrighted works. But to facilitate more rapid shipment of books to India (to "feed" the scanners), plans evolved to allow for collections of books to be acquired and shipped to India without knowing their copyright status. The University Libraries needed some way to discover what copyrighted books had been scanned, and the delivery system needed some way to "know" whether permission was granted to include the copyrighted books in the Million Book Collection. Conceived originally to track copyright and permission status, the administrative metadata database quickly expanded to include additional data needed to manage the Project. For example, in addition to copyright and permission status, the database will also track the source library that provided the books, whether the books need to be returned to the source library, [10] and whether digital copies of the books and associated bibliographic metadata and OCR have been provided to the publishers, the contributing libraries, and all mirror sites.

A prototype of the administrative metadata database will be available and introduced to our Indian partners when they visit Carnegie Mellon in May 2003. Capturing metadata on copyright[11] and permission status, the source library, and whether the books need to be returned to the source library, will be added to the workflow in India. Indian partners will generate reports per library of titles that need to be returned to ensure that the appropriate books are returned to the right library. The University Libraries will generate reports of in-copyright books with permission status unknown, and then consult the publisher database to see whether negotiations are already underway with those publishers. If not, we'll begin the process of seeking copyright permission. Project partners in the United States will update permission status in the administrative and publisher databases when negotiations are complete and permission has been granted or denied. Updates of the administrative metadata will be automated so that the records of all titles for which a publisher has granted or denied permission can be updated at once. Copyrighted works will not be available in the Million Book Collection unless or until the administrative metadata indicates that permission has been granted. The University Libraries will also generate reports and update the administrative database to ensure that publishers, contributing libraries, and mirror sites receive appropriate copies of the digitized books.

Data in the publisher and administrative metadata databases will be used to conduct quantitative assessments of the copyright permission work. Plans also include surveying participating publishers to determine their satisfaction with the quality of the copies they received, what they did or plan to do with the copies, and the impact, if any, that participation in the Million Book Project had on their revenue and their view of open access to copyrighted works.

Looking Ahead

Learning how to reduce the cost of seeking copyright permission and how to negotiate successfully with copyright holders has become essential for balancing the public's right to access to ideas with the copyright holder's private interest in generating revenue from their work. Carnegie Mellon University Libraries will share the publisher database with other groups involved in seeking copyright permission to digitize academic materials, and disseminate lessons learned through negotiations with publishers. Publisher contact information and tested and

[10] To reduce the cost of shipping, we're trying to acquire duplicate or weeded titles that do not need to be returned to the source library.

[11] Determining copyright status entails consulting the copyright renewal records for books published 1924-1964.

documented negotiation strategies should reduce the cost of seeking copyright permission in the future. If successful, the Million Book Project could turn the tide in whether our cultural and intellectual heritage turns to dust or digital.

Bibliography

Friedlander, Amy. Dimensions and Use of the Scholarly Information Environment: Introduction to a Data Set. Washington DC: Council on Library and Information Resources publication 100, October 2002. Available: <www.clir.org/pubs/reports/pub110/contents.html>

Guthrie, Kevin M. *Revitalizing Older Published Literature: Preliminary Lessons from the Use of JSTOR* (March 23, 2000). Available: <http://www.jstor.org/about/preliminarylessons.html>

Kline Pope, Barbara. “National Academy Press: A Case Study.” *The Journal of Electronic Publishing* 4, 4 (June, 1999). Available: <http://www.press.umich.edu/jep/04-04/pope.html>

Lawrence, Steve, and Lee Giles. “Accessibility and Distribution of Information on the Web.” *Nature* 400 (1999): 107-109. Summary of findings available: <http://www.wwwmetrics.com>

Mardis, Marcia. “Uncovering the Hidden Web, Part I: Finding What the Search Engines Don’t.” *ERIC Digest* EDO-IR-2001-02 (October 2001). Available: <http://www.ericit.org/digests/EDO-IR-2001-02.shtml>

OCLC. *How Academic Librarians Can Influence Students’ Web-Based Information Choices*. (OCLC White Paper on the Information Habits of College Students.) Dublin, Ohio: OCLC, June 2002. Available: <http://www2.oclc.org/oclc/pdf/printondemand/informationhabits.pdf>

Quint, Barbara. “Overture Acquires Two Major Web Search Engines.” *Information Today, Inc. Online* (March 3, 2003). Available: <http://www.infotoday.com/newsbreaks/nb030303-1.shtml>

Ringer, Barbara A. “Study No. 31: Renewal of Copyright.” In *1 Studies on Copyright*, Arthur Fisher Memorial ed., June 1960: 513-514.

searchWebServices.com Definitions. See “deep Web.” http://searchwebservices.techtarget.com/sDefinition/0,,sid26_gci558034,00.html dated June 5, 2001.