

*Penn Libraries*

*Scholarship at Penn Libraries*

---

*University of Pennsylvania*

*Year 2007*

---

Copyright and Provenance: Some  
Practical Problems

John Mark Ockerbloom  
University of Pennsylvania, ockerblo@pobox.upenn.edu

In Bulletin of the IEEE Computer Society Technical Committee on Data Engineering,  
Vol. 30, No. 4, Dec. 2007, pp. 51-58

This paper is posted at ScholarlyCommons.  
[http://repository.upenn.edu/library\\_papers/47](http://repository.upenn.edu/library_papers/47)

# Copyright and Provenance: Some Practical Problems

John Mark Ockerbloom  
University of Pennsylvania  
ockerblo@pobox.upenn.edu

## Abstract

*Copyright clearance is an increasingly complex and expensive impediment to the digitization and reuse of information. Clearing copyright issues in a reliable and cost-effective manner for works created in the last 100 years can involve establishing complex provenance chains for the works, their copyrights, and their licenses. This paper gives an overview of some of the practical provenance-related issues and challenges in clearing copyrights at large scale, and discusses efforts to more efficiently gather and share information and its copyright provenance.*

## 1 Introduction

As information seekers increasingly move from print to digital media, print resources are being digitized at an ever-accelerating rate. As of 2007, over a million volumes have been digitized by libraries such as the Library of Congress and the University of Michigan, for-profit corporations like Google, and public-private partnerships such as the Open Content Alliance [1]. Mass digitization is made possible by ever-lower costs for large-scale scanning and storage. The Open Content Alliance's scanning projects for example, digitize books nondestructively at a cost of 10 cents per page, or about \$30 for a 300-page book [2].

The cost of clearing copyright, however, can be substantially higher than the cost of digitization itself. A 2003 study of attempts to obtain copyright permissions for a book digitization project at Carnegie Mellon University found that it cost \$78 per title to clear copyrights of the copyrighted books they sought to digitize [3]. This figure does not include any royalty costs, but only the overhead cost in determining copyright status and obtaining necessary permissions. Most of this cost was labor, a cost that tends to increase over time.

Most books, particularly those by a single author, have relatively few and simple copyrights. However, periodicals, collective works, sound recordings, and motion pictures often involve a large number of potential copyrights and copyright owners in their various elements. Moreover, different rights and permissions may apply to these copyrights in different contexts and legal jurisdictions.

There is widespread scientific, business, and cultural interest in disseminating, adapting, and reusing the content of others, as seen in initiatives like the Internet Archive, Google Book Search, YouTube, and Arxiv.org. Since copyright restrictions apply to most present-day content, as well as historic content going as far back as 100 years, clearing copyright can be both an essential and a costly part of these initiatives.

In order to legally publish and reuse content, one typically needs to determine the answers to several important questions:

---

*Copyright 2007 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.*

**Bulletin of the IEEE Computer Society Technical Committee on Data Engineering**

---

- What copyrights, if any, apply to this content? (And which are currently in force?)
- Who controls these copyrights, and how can they be contacted?
- What permissions, if any, have been granted concerning the copyrighted content?

In projects that make large-scale reuse of content, it may not be feasible to have complete, certain answers to these questions for all the content one may want to use. In practice, content reuse at scale may be best understood as an optimization problem along multiple dimensions, according to these desires:

- **Maximizing the value of the collection**, by including as many valuable resources as one can, with the broadest rights possible.
- **Maximizing throughput on rights-clearing**, so a large collection can be built quickly.
- **Minimizing the cost of rights-clearing**, which as noted above can be impractically high per item.
- **Minimizing the risk of legal penalties**, which in the worst case can be very large. Current copyright law in the US authorizes statutory penalties (which are distinct from penalties for actual damages) of up to \$150,000 per infringement. Penalties are lower if the infringement is shown not to be willful, but proving that in court can be uncertain and costly, and even non-willful infringement statutory penalties can run into the thousands of dollars [4].

Different projects may put different priorities on these dimensions. For example, Google’s Library project has to date been especially conservative with copyright determination in its “full view” book displays. In some cases it presumes that US copyright is still in force for books published as long ago as 1909, and that foreign copyrights may subsist from as far back as 1865. (In a large number of these cases, these copyrights have in fact long since expired.) These conservative guidelines, intended to minimize clearance cost and risk, enable them to make visible a significant number of books with very little effort needed to clear titles, but suppresses much content that could be usefully viewed and repurposed by the public. In contrast, organizations like the Internet Archive put a higher priority on maximizing collection value in the copyright clearance of their digitized texts, and therefore expose many commercially published works from as late as 1963, and government-published works up to the present day.

In order to control risk and cost of clearing copyright in a large-scale project, simply stating copyright restrictions in binary terms (such as “this is in the public domain” or “this is in copyright” is insufficient. To understand the reliability and applicability of such determinations, one needs to know a variety of facts that inform those determinations, and know how these facts were derived. Moreover, the same facts may lead to different determinations in different places, times, and contexts. What may be legal to reuse in a classroom in the US in 2007 may not be legal to reuse in a commercial film in Japan in 2009, or vice versa, but the determination in both of these contexts may depend on the same underlying set of facts about the work in question and its copyrights.

Managing these facts involves several kinds of provenance issues. To reliably determine the rights to a work, one may have to understand and record the provenance of a work, the provenance of its rights, and the provenance of the information used in rights determination. In the next sections, I survey some of the specific provenance problems involved, and describe some of the derivations and uncertainties that are part of copyright clearance. I then describe some methods to alleviate the problems of provenance in copyright determinations, and suggest ways in which copyright clearance can be a productive and illuminating application domain for provenance research.

## 2 Provenance issues

In this section, I describe some of the provenance-related issues involved in determining a work's copyright status and the rights available for using the work. It is not within the scope of this paper to provide a complete or authoritative guide to copyright clearance; rather, I illustrate important aspects of clearance where provenance is relevant. Useful detailed guides to copyright clearance in the US, written by copyright attorneys, include [5] and [6].

### 2.1 Provenance of works

Whether a work is copyrighted at all, what copyrights might apply to it, and who initially owns the copyrights to the work, depend on the provenance of the work itself. Relevant questions include:

- **Who authored the work?** And when did they live? Copyrights are generally assigned to authors by default, and in many cases last for a set period after the author's death.
- **When was the work created, first published, first published with a US-recognized copyright notice, and first published in the US?** In various cases, the time and place of these events determine when a copyright term starts and ends.
- **Is this a work for hire? For whom?** Works for hire may have different initial copyright owners and copyright term lengths. Works produced as work for hire for the US government may not be copyrighted at all.
- **Does this work include or derive from other works?** If so, the rights available for this work may depend on the rights available for those other works.
- **What is the work's current commercial status?** For example, if a book is sufficiently old, out of print, and cannot be bought inexpensively, some US users may have special rights to reuse the work without permission under special provisions of US copyright law, even if the work is still under copyright [7].

### 2.2 Provenance of rights

The rights available for the use of a work depend on a number of factors apart from the provenance of the work itself. The provenance of the rights must also be considered. In some jurisdictions, for instance, copyrights must be explicitly asserted and maintained through various mechanisms in order to remain in force. Rights to a work can also be transferred, in whole or in part, from the original author to new agents. (This is commonly done with articles submitted to scholarly journals, for example.) The original or subsequent rightsholders can further license the work under various terms and conditions. These transfers and licenses may be matters of public record, or private agreement.

Under the Berne Convention, the dominant international copyright treaty, copyright automatically applies to a new creation without any formal claims or registration. While this principle may make it easier to determine that a copyright has not prematurely expired, it may make it more difficult to determine when the copyright was originally established, and who claimed it.

Many copyrights today, however, were originally established under different regimes than that of Berne. For example, the United States was a latecomer to the Berne treaty and until the early 1990s made copyright status dependent on notices, registration, and formal renewal of copyright. Under US law, valid copyright notices include an explicit claim of copyright, a year in which the copyright was claimed, and the name of the claimant. Registration and renewal involves providing certain data about the work, including the author, title, and date of claim. The US Copyright Office records, maintains and makes available the data in these registrations.

Relevant questions of rights provenance, then, include:

- **What copyright notices, if any, were distributed with the work?** These notices may determine the copyright status of a work, as well as noting the initial owner and the start of the work’s copyright term.
- **What registrations and renewals were made of the copyright?** These may also determine the copyright status of a work and identify owners. Renewals may show ownership changes since the initial registration.
- **What assignments were made of the copyright or of subsidiary rights?** These may involve full copyright transfers, or narrower assignment of rights for certain uses and jurisdictions. For example, a freelance writer may assign US first serial rights to an article to a particular magazine, but retain rights to republish the work in other formats and markets. The exact terms of such rights assignments are typically governed by contract language rather than by statute, though in some cases, such as intestate or insolvent authors, rights assignments may be determined by local inheritance or bankruptcy laws. Rights may be assigned to a specific party, or in some cases to the world at large. Open source and Creative Commons licenses, for example, specify that anyone has certain rights to use a work under standard terms and conditions that are typically published along with the work.

### 2.3 Provenance of information

While provenance of the work and the rights are sufficient in theory to determine whether and how a work can be used, in practice one cannot rely on perfect and complete knowledge of this information. Therefore, those who wish to make copyright determinations must also consider the provenance of the information they have about the works and copyrights. What are the sources of information? Are they reliable? Did they derive their information from other sources? If so, which ones? Are there important sources of information that are not being taken into account, and could these change one’s copyright determinations in important ways?

Many rights determination issues include or derive from negative as well as positive information. For example, consider the judgment that a book first published in 1940 in the US is in the public domain in that country. Positive information supporting this judgment may include the imprint of a US publisher, and the notations “first edition” and “copyright 1940” on the title and verso pages of the book. Negative information may include the lack of any prior editions of the book, the lack of any further copyright notices in the book, the lack of a copyright renewal, and the lack of any prior publications from which the book derives.

Some of these facts may be easier to establish than others, with negative information usually more difficult to prove than positive information. The imprint and copyright notice of a book, for instance, can be verified with images of the pages on which they appear. The lack of other copyright notices might be established from other pages on which copyright notices might appear, which can be a larger page set. The lack of copyright renewal can be verified against a complete data source of copyright renewals, which exists, but which is in turn a much larger information set that is more difficult to access or search in full than the information sets discussed to this point. The lack of previous editions or works from which the book might derive depends on yet larger, and less well-defined, information domains. In practice, at some point along the continuum of verifying these facts, one will need to rely on the judgment of another person or source, rather than including the complete set of information needed to establish a particular fact.

The particular source of this information is important. One might trust the word of a publisher or a professional librarian about the copyright date or status of a book more than the word of an anonymous uploader to a file-sharing site. However, sometimes unexpected information can surprise even experts. In 2004, a popular online animated political satire used, without permission, the tune and some of the words of Woody Guthrie’s song “This Land is Your Land” The animators were sent a cease and desist notice by Guthrie’s music publishers, who had duly registered and renewed the copyright on their first edition of the work. A complaint brought by the Electronic Frontier Foundation on behalf of the producers of the animation initially relied on a fair use defense; however, in the course of litigation, the Foundation discovered that Guthrie had produced a hand-written song book, complete with copyright notice and cover price, that included an early version of the song, years before the

song was conventionally published. Only a few known copies of this work are known to exist, and its copyright was never renewed. The publisher and the animators settled quickly thereafter, with an agreement allowing the online animation to continue [8].

In some cases, positive or negative information asserted about the copyright of a work may need to be discounted or overridden. It is not unheard of, for example, for a publisher to place a copyright notice, dated the year of publication, on an unaltered reprint of a public domain work, even though under US law such reprints are not entitled to a new copyright. Contributors to shared content sites like YouTube or Wikipedia may attach liberal licenses to content, authored by others, that they do not have the right to relicense or redistribute. For some uses, such as the literal reproduction of works that carry copyright notices, it may be important to retain these assertions while at the same time noting that they do not apply, or do not apply in full.

### **3 Derivations and uncertainties**

How far back one needs to trace and record the provenance of copyright information depends on the relative importance of risk, cost, and productivity in the optimization problem described earlier. Their importance, and the degree of copyright provenance recording required, may vary depending on context and application. Project Gutenberg, for instance, requires and stores title and verso page images to establish original claims of copyright. It also has produced, and uses, a text transcription of the book renewal sections of the US Copyright Office's Catalog of Copyright Entries, in order to determine whether a book copyright has been renewed. For other fact determinations relevant to copyright, however, such as the lack of prior publications of the work, it relies on the judgment and assertions of its contributors and volunteers.

The transcription of the Catalog of Copyright Entries itself is derived from earlier artifacts. The ultimate source of a copyright renewal claim is the renewal form filed by a copyright holder and deposited with the US Copyright Office. These forms are included or reproduced in registration books that are accessible to Copyright Office staff. From 1978 onward, the information in these forms has been used to populate an online database accessible worldwide. Before 1978, though, catalog cards were prepared from these forms to allow copyright registrations and renewals to be looked up by name, title, or various other criteria. These cards are accessible to both the Copyright Office staff and to members of the general public that can visit the copyright card catalog in Washington, DC. From these cards, bound volumes of the Catalog of Copyright Entries were printed and distributed to libraries across the United States and beyond, where they are available to patrons of those libraries (though in many libraries the volumes are kept in closed reserve stacks). From this point, already some distance down the provenance chain, independent third parties have made digital images of some of the Catalog of Copyright Entries pages and published them on the Internet, where the digital images have been used to produce text transcriptions of the copyright registration information that are used by Project Gutenberg and others.

The renewal records that can be quickly searched online, then, can be information derived via several steps from the original copyright holder filings. It is possible that errors or omissions exist in the derivations, and that erroneous rights determinations may be made as a result. Going back further brings one closer to the original copyright filings, but is generally more difficult and costly. As mentioned above, the Catalog of Copyright Entries have been partially digitized. The copyright card catalog has not, nor have the registration books. Digitizing these resources would be more expensive than digitizing the summary Catalog, due to the larger number of images and the rarity and vulnerability of the materials. They would also be more cumbersome to search than a database would be. Thus, we see that following provenance chains further back involves a tradeoff of cost and risk. Studies at Stanford suggest that the error rate of using transcriptions of the Catalog of Copyright Entries is very low, and that errors in derivation were much less common than errors and term mismatches in copyright searches [9]. Many projects, then, may well find the more easily searched derivative forms of the copyright records a useful or even superior starting point for research.

In many cases, simply searching copyright records will not answer the question of what can be done with

a work. If the records indicate that a work is still under copyright, or there is not sufficient information to determine with sufficient certainty that a work is no longer under copyright, then one may be legally liable if one reuses the work. To avoid this liability, one must obtain permission (or assurances of public domain status) from the presumed rightsholder. Unfortunately, for many works the rightsholder can be difficult or impossible to determine or contact. Copyright claimants at the time of registration may be listed in the Catalog of Copyright Entries, but their addresses are not (though they may appear in the registration books). Copyrights may have been transferred, assigned or willed to others since the copyright was registered or renewed, and this is more likely the longer that the copyright term has run. While copyright transfers may be registered with the Copyright Office, there is no requirement that transfers be registered there or anywhere else.

Hence, there is now a large and growing set of “orphan works” for which copyright cannot be reliably cleared, due to the inability to determine or locate the proper copyright owners. Orphan works come in all varieties: “abandonware” developed by defunct software companies; articles and monographs from long-dead authors with obscure heirs; images of great historic or artistic importance whose original creator cannot be traced; documentary productions and compilations that feature copyrighted material from a wide variety of creators, not all of whom can be determined or found. As US law stands now, orphan works are effectively impossible to reuse legally (beyond the usual rights of fair use and resale). The Copyright Office has acknowledged the orphan works problem as a serious one, and has held public hearings and suggested legislation to alleviate it.

Uncertainties can also exist with rights to data. In many European countries, factual data can have copyright-like restrictions associated with it. (In the US, facts in themselves are in the public domain, though an original selection, expression, or arrangement of the facts can be copyrighted.) Also, in many legal jurisdictions, privacy laws or confidentiality agreements may limit the disclosure of certain data. Keeping track of rights and restrictions on private information, while often not specifically a copyright issue, involves many of the same issues of provenance tracking as copyright clearance does.

## **4 Initiatives for easing copyright clearance**

Copyright clearance need not be as complicated and risky as it currently is. Several initiatives have been started or proposed to ease copyright clearance, a number of which relate to provenance.

One way to ease copyright clearance at the large scale is simply to promulgate standards for recording and distributing copyright-related information. For example, many of the digitized books at the Internet Archive include metadata relevant to copyright in standard forms, including publication dates, copyright notice information, and the results of copyright renewal searches. While this information is not currently formatted for machine processing, its use of a standard vocabulary and set of assertions about copyright has inspired efforts to define standard, structured, machine-readable vocabularies and grammars for expressing copyright facts [10]. Note that the vocabulary of copyright provenance is different from the vocabulary of digital restrictions often used by Digital Rights Management (DRM) systems. The latter is largely concerned with the specific operations that software should allow or disallow on particular content, such as printing, reading aloud, or duplicating. The former is concerned with underlying intellectual rights and permissions, such as the existence and owner of copyright, the duration of the copyright, and licenses granted for the content. DRM restrictions may be derived in part from these underlying rights, but are distinct from them.

One vocabulary of rights expression that has gained widespread popularity in recent years is the Creative Commons vocabulary. Creative Commons supports a variety of standardized permissions, such as the right to reuse with attribution, or noncommercially, or without making derivatives, or with the right to make derivatives that must be licensed under the same terms as the original. These permissions can be encoded in machine-readable format and distributed along with, or as part of, a copyrighted work. Assuming that the permissions were granted by an authorized party, this rights expression system allows others to easily reuse a work in well-understood ways, without having to trace the copyright holder to get special permissions [11].

Registries are useful as stores of copyright clearance data, including provenance data. The US Copyright Office is one such registry already discussed, but other types of registries also exist or have been proposed. For example, the Writers, Artists, and Their Copyright Holders registry, based in the US and the UK, keeps up-to-date contact information on many well-known copyright holders [12]. Groups representing copyright holders, such as the Copyright Clearance Center and the Harry Fox music licensing agency, both track copyright holders of works and handle payments to them, streamlining common uses of many copyrighted works such as recording new performances of songs or reprinting articles from periodicals. The Online Computer Library Center (OCLC) has proposed a general purpose registry of copyright information to be associated with its WorldCat union catalog, to make it easier to clear rights for all kinds of library materials [13].

Legislation can also ease copyright clearance, and suggest particular types of provenance information that may be useful to track. For example, the proposed Orphan Works Act of 2006 would have allowed copyrighted works to be reused by others without permission if the users were unable to find the copyright holder after a “reasonably diligent search” [14]. To date, orphan works legislation has not been enacted in the US, but if it were, it could limit the degree of provenance information one would need to gather for a work (since one might not have to trace back copyright information indefinitely if doing so were unreasonably burdensome). On the other hand, to prove that a reasonably diligent search was conducted, users might wish to explicitly document the steps taken in the search, to establish the provenance of one’s “reasonably diligent” determination.

## 5 Conclusions

The preceding discussion should illustrate how provenance issues are important in copyright clearance, and how copyright clearance is a significant application domain for provenance research. Practical, reliable copyright clearance requires careful consideration of the provenance of works, their rights, and the assertions about the works and rights. Optimal procedures for rights determination must take into account positive and negative factual assertions, legal analysis tailored to jurisdiction and context, and an appropriate balancing of value, throughput, cost and risk. The factual assertion chains that support determinations of rights available for works can be complex in their structure, and involve varying degrees of uncertainty.

Copyright clearance, then, is fertile ground for applying provenance research. Theoretical foundations for evaluating the reliability of assertion chains can be applied to estimate risk in copyright determinations. Common representations of copyright assertions, searches, and their provenance, can be preserved as metadata and used in context-sensitive copyright evaluations. Simple, cheap methods of storing and querying this provenance information, and improvements in the efficiency of provenance calculations, data representations, and queries, can improve the reliability and practicality of copyright evaluations in large-scale collections.

Moreover, improved copyright clearance is not simply an interesting research application. The easier it is to safely and legally reuse the works of the past, the easier it becomes to advance the state of knowledge and culture. The technologies that now allow organizations to digitize millions of books for the Internet make it possible to revive, redistribute and build upon the large corpuses of text, data, audiovisual media, and software, that make up the historic, cultural, and scientific endowment of the world. If advances in provenance handling allow us to more easily clear their copyrights, we may all enjoy greater access to a richer heritage of knowledge. As Isaac Newton and other scientists have noted, building on this richer heritage can let us all see farther, standing on the shoulders of giants [15].

## References

- [1] Sharita Forrest, “An Open Book: CIC Member Libraries Join Google in Digitizing up to 10 Million Volumes” *Inside Illinois*, 26:22, June 2007, online at <http://www.news.uiuc.edu/ii/07/0621/google.html>.
- [2] “Open Content Alliance Will Scan Boston’s Books” *Chronicle of Higher Education*, September 28, 2007; online at <http://chronicle.com/wiredcampus/article/2418/open-content-alliance-will-scan-the-best-of-bostons-books> .
- [3] Denise Troll Covey, *Acquiring Copyright Permission to Digitize and Provide Open Access to Books*. Digital Library Federation and Council on Library and Information Resources, October 2005. Online at <http://purl.oclc.org/dlf/pubs/dlf105/>
- [4] Statutory penalties are specified in section 504 of the Copyright Act, “Remedies for Infringement: Damages and Profits” (17 USC 504). Online at <http://www.copyright.gov/title17/92chap5.html#504> .
- [5] Stephen Fishman, *The Public Domain: How to Find Copyright-Free Writings, Music, Art and More*. (3rd edition; Berkeley, CA: Nolo Press, 2006.)
- [6] William S. Strong, *The Copyright Book: A Practical Guide*. (5th edition; Cambridge, MA: MIT Press, 1999.)
- [7] See section 108 of the Copyright Act, “Limitations on exclusive rights: Reproduction by libraries and archives” (17 USC 108), online at <http://www.copyright.gov/title17/92chap1.html#108> . The right of libraries to reproduce older works no longer commercially exploited under certain conditions is spelled out in section 108(h).
- [8] Katie Dean, “JibJab is Free for You and Me” *Wired News*, August 24, 2004; online at <http://www.wired.com/entertainment/music/news/2004/08/64704> .
- [9] Mimi Calter, “Assessing Copyright Status: Copyright Renewals Database” Presentation at Digital Library Federation Forum, November 2007. Online at <http://www.diglib.org/forums/fall2007/presentations/Calter.pdf> .
- [10] “Data elements needed to ascertain copyright facts” MARC Discussion Paper No. 2007-DP05, May 30, 2007. Online at <http://www.loc.gov/marc/marbi/2007/2007-dp05-original.html> .
- [11] Creative Commons website. Online at <http://creativecommons.org/> .
- [12] The Watch File: Writers, Artists and Their Copyright Holders, online at <http://tyler.hrc.utexas.edu/> .
- [13] See comment by Bill Carney of OCLC in O’Reilly Radar, November 7, 2007, online at [http://radar.oreilly.com/archives/2007/11/checking\\_copyri.html](http://radar.oreilly.com/archives/2007/11/checking_copyri.html) .
- [14] Gigi Sohn, “Orphan Works Bill Introduced” *Public Knowledge Policy Blog*, May 22, 2006. Online at <http://www.publicknowledge.org/node/392> .
- [15] Newton’s famous aphorism, used by numerous writers both before and after his time, is discussed in great detail in Robert K. Merton’s book *On the Shoulders of Giants: A Shandean Postscript* (Chicago: University of Chicago Press, 1993). I verified the aphorism, and Newton’s use of it, with a searchable digital copy of the book provided by Amazon at <http://www.amazon.com/exec/obidos/ASIN/0226520862> .