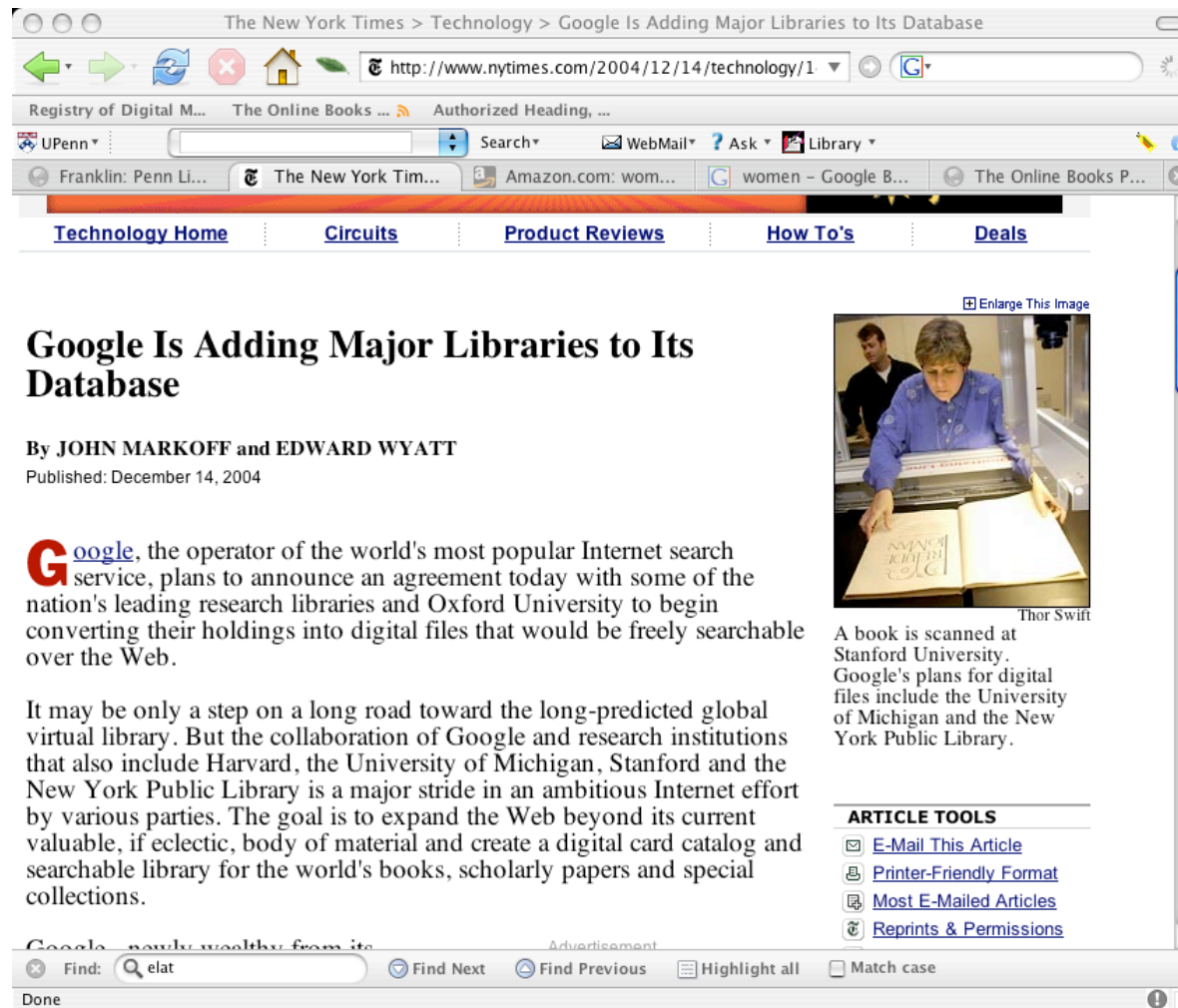


Copyright and provenance

Some practical problems

John Mark Ockerbloom
Principles of Provenance workshop
June 26, 2007

Our resources are going digital



The screenshot shows a web browser window with the URL <http://www.nytimes.com/2004/12/14/technology/1>. The page title is "The New York Times > Technology > Google Is Adding Major Libraries to Its Database". The browser's address bar shows the URL. The page content includes a navigation menu with links for "Technology Home", "Circuits", "Product Reviews", "How To's", and "Deals". The main article title is "Google Is Adding Major Libraries to Its Database" by JOHN MARKOFF and EDWARD WYATT, published on December 14, 2004. The article text discusses Google's plans to digitize books from major libraries. A photograph shows a woman scanning a book. The article tools section includes links for "E-Mail This Article", "Printer-Friendly Format", "Most E-Mailed Articles", and "Reprints & Permissions".

The New York Times > Technology > Google Is Adding Major Libraries to Its Database

Registry of Digital M... The Online Books ... Authorized Heading, ...

UPenn ... Search ... WebMail ... Ask ... Library ...

Franklin: Penn Li... The New York Tim... Amazon.com: wom... women - Google B... The Online Books P...

[Technology Home](#) | [Circuits](#) | [Product Reviews](#) | [How To's](#) | [Deals](#)


Google Is Adding Major Libraries to Its Database

By JOHN MARKOFF and EDWARD WYATT
Published: December 14, 2004

Google, the operator of the world's most popular Internet search service, plans to announce an agreement today with some of the nation's leading research libraries and Oxford University to begin converting their holdings into digital files that would be freely searchable over the Web.

It may be only a step on a long road toward the long-predicted global virtual library. But the collaboration of Google and research institutions that also include Harvard, the University of Michigan, Stanford and the New York Public Library is a major stride in an ambitious Internet effort by various parties. The goal is to expand the Web beyond its current valuable, if eclectic, body of material and create a digital card catalog and searchable library for the world's books, scholarly papers and special collections.

[Enlarge This Image](#)



Thor Swift

A book is scanned at Stanford University. Google's plans for digital files include the University of Michigan and the New York Public Library.

ARTICLE TOOLS

- [E-Mail This Article](#)
- [Printer-Friendly Format](#)
- [Most E-Mailed Articles](#)
- [Reprints & Permissions](#)

Find: elat Find Next Find Previous Highlight all Match case

Done

But not without a fight

Publishers sue Google over book search project

By Alorie Gilbert

Staff Writer, CNET News.com



Published: October 19, 2005, 9:37 AM PDT

 [TalkBack](#)  [E-mail](#)  [Print](#)  [del.icio.us](#)  [Digg this](#)

Welcome Google User!

More headlines related to: "copyright books lawsuit millions":

- [Fighting to protect copyright 'orphans'](#)
- [Amazon files objection to Google subpoena](#)
- [Copyright tussles for Google](#)
- [More matching headlines >](#)

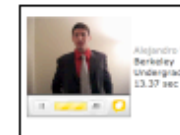
 [Add News.com to Google](#)  [Add to Google](#) Add CNET News.com headlines to your Google homepage or Google reader.

The Association of American Publishers has filed a lawsuit against Google, alleging the Internet company's plans to scan and digitally distribute the text of major library collections would violate copyright protections.

The group filed suit after lengthy discussions with Google's management about the company's [Print Library Project](#) broke down, the AAP said on Wednesday.

As part of the project, Google is working to scan all or parts of the book collections of the University of Michigan, Harvard University, Stanford

▼ adve



Featured gallery
Video: Getting
Alumwire co-found with CNET News.co inspired their start-recent college grac

What can I do with a work?

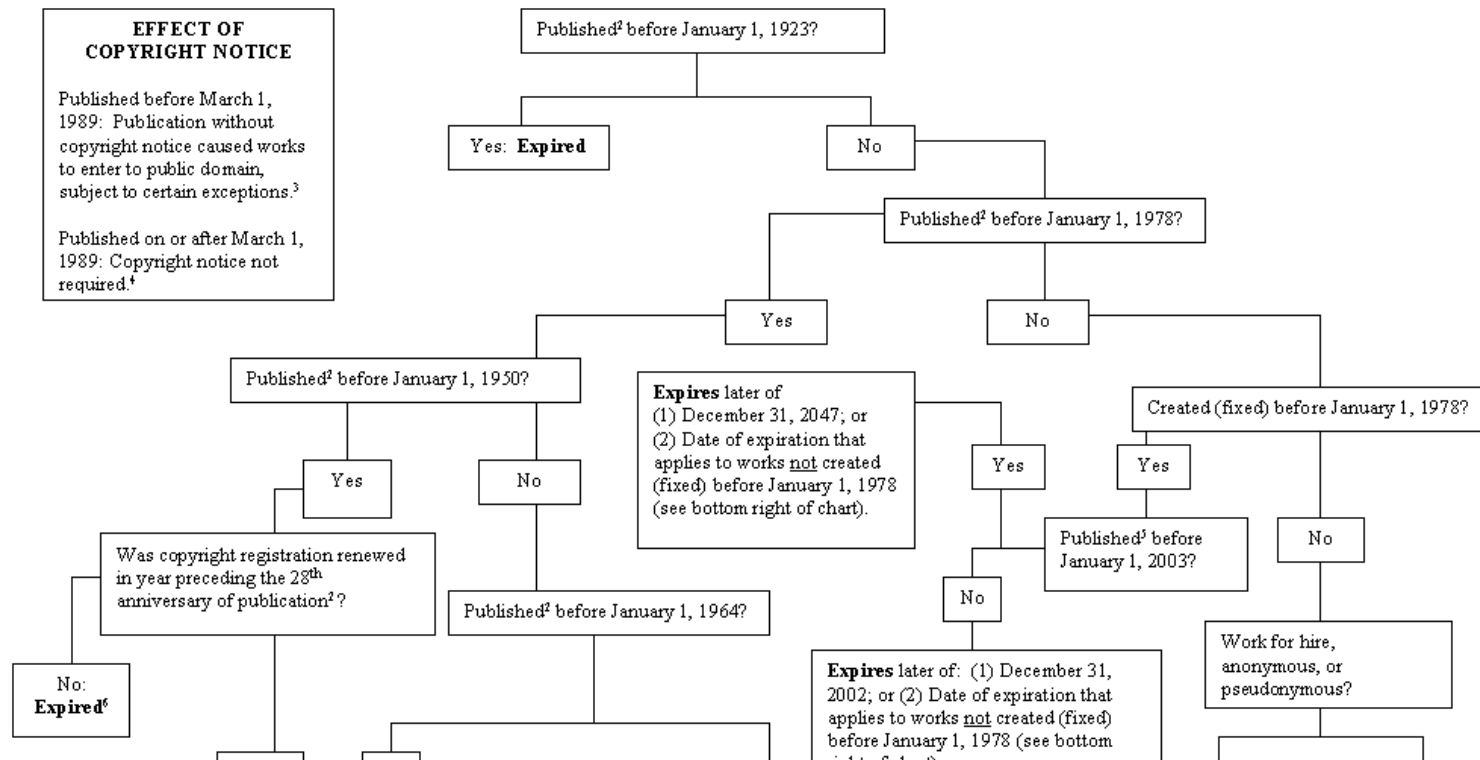
- **Often simply represented as a binary value “restricted access” / “no restrictions”**
- **But really derives from a series of copyright questions**
 - **What copyrights apply to this work?**
 - » **Which are still in force? Which have expired?**
 - **Who holds the copyrights?**
 - » **And how can they be contacted?**
 - **What permissions apply for use?**
 - » **To specific parties (first serial rights to a magazine..)**
 - » **General (GPL, Creative Commons...)**
 - » **Granted by law (compulsory license, fair use, preservation, orphan works...)**

Google's conservative algorithm for showing more than snippets

```
Boolean can_show_full_without_permission
(client, publication_place, publication_date) {
  If (client in US) {
    If ((publication_place in US) {
      return true if (publication_date < 1923);
    } else {
      return true if (publication_date < 1909);
    }
  } else {
    return true if (publication_date < 1864);
  }
  return false;
}
```

A portion of a copyright expiration flowchart

FLOWCHART FOR DETERMINING WHEN U.S. COPYRIGHTS IN FIXED¹ WORKS EXPIRE



Costs to produce digital works

- **Digitization costs**
 - \$0.01 - \$0.10/page at scale
 - \$5-\$50 for a 500-page book
- **Copyright clearance costs**
 - 2003 Carnegie Mellon study: \$78 on average per book not in “conservative algorithm” territory
 - Books are relatively simple; magazines, recorded music, film, typically involve lots of rightsholders

An optimization problem in multiple dimensions

- **Maximize value of collection**
 - (include as many valuable resources as you can, with broadest rights possible)
- **Maximize throughput of rights-clearing**
 - (to build a large collection fast)
- **Minimize cost of rights-clearing**
- **Minimize risk of legal penalties**
 - (which can be very large in worst case)

Some provenance questions in copyright determination

- **Provenance of work**
 - Who created the work? (And when did they live?)
 - When and where was the work first published?
 - » (inside and outside the US)
 - Does the work include or derive from other copyrighted works? Which ones?
- **Provenance of rights**
 - Was there a copyright notice? What did it say?
 - Is this a work for hire? For whom?
 - Was the work registered for copyright? When? Renewed?
 - Was copyright transferred? If so, when and to whom?
 - Record of rights permission grants?
- **Determination can work from copyright knowledge bases or from works, involve negative as well as positive assertions with various degrees of certainty**

An example



Copyright registrations

- **Rightsholders filled out registration forms**
- **Form copies in registration books in DC**
 - The earliest saved artifact
- **Form data summarized in card files in DC**
 - What Copyright Office staff search
- **Card files summarized in Catalog of Copyright Entries in various libraries**
 - What folks outside DC tend to search
- **Digitizations, databases produced**
 - Page Image index at Penn; Stanford's Determinator database
- **Digitization cost depends on where you start**

Some recent developments

- **OCLC proposed project to associate copyright-related assertions with WorldCat entries in global union catalog**
- **Proposals to exempt “orphan works” from infringement liabilities**
 - **But need to conduct “reasonably diligent search”; and document that the search was done**

What do we need, and how can provenance research help?

- **Theoretical foundations for evaluating sufficiently reliable information chains**
 - What facts support our copyright determinations, and what is their provenance?
- **Practical issues of scaling up evaluation**
 - (preferably with automation)
- **Common representation for copyright assertions and searches, and their provenance**
 - Accommodating distributed contributions, evaluation
- **Simple, cheap methods of storing and querying this information**