April 11, 2006

The Next Mother Lode for Large-scale Digitization? Historic Serials, Copyrights, and Shared Knowledge

John Mark Ockerbloom, University of Pennsylvania

Available at: https://works.bepress.com/john_mark_ockerbloom/5/
The next mother lode for large-scale digitization?

Historic serials, copyrights, and shared knowledge

John Mark Ockerbloom
DLF Spring Forum, Austin
April 11, 2006
Main ideas

• A vast amount of significant serial literature before 1964 is in the public domain in the US
  – Both scholarly and general-interest content
  – More complete, and potentially more accessible, view of mid-20th-century culture and thought than public domain books

• We can determine what is available to digitize
  – We have created an inventory of all periodicals renewals 1950-1977 (for 1923-1950 publications; only a tiny fraction renewed)
  – This inventory can be the germ of a more comprehensive, cooperatively built knowledge base

• Leadership opportunities for DLF and its institutions
  – We have the big serial collections, the hard-core users, the knowledge of the literature and of digital library issues
  – Low-overhead shared knowledge bases can provided a basis for coordinating work
The hopes of mass digitization

“It is time to build… an America where every child can stretch a hand across a keyboard and reach every book ever written, every painting ever painted, every symphony ever composed…”

-- US President Bill Clinton, 1998 State of the Union address

“Our venture will result in a magnitude of discovery that seems almost incomprehensible…”

-- U-Michigan President Mary Sue Coleman, 2006 speech to Association of American Publishers, on the Google Book Search project
What can we actually make available online?

• Every book, painting, symphony…
  – That is in the public domain
  – Or that we can license (actively or passively)
  – Or that we can get a special exemption to provide
    » fair use
    » section 108
    » orphan works?
    » Snippets of fully indexed text?
What are we getting from mass text digitization efforts so far?

• Many books before 1923, or 1909, or 1864
  – Over 40,000 from Google alone since the fall
  – Public domain status determination often very conservative

• Some open access serial runs
  – Open access largely seen in nonprofit projects (MOA, newspapers, ILEJ…) or by journals themselves (smaller-scale)
  – Google, Internet Archive, etc. also including some serial volumes along with the books, not systematically to date
  – Larger collections available in limited access (EBSCO, JSTOR…)

• A fascinating assortment of miscellaneous text
  – Pamphlets, manuscripts, letters, diaries, blogs, ephemera…
  – We won’t focus on that in this talk
What do we get from all this text?

- Classic or timely wisdom and art
  - Would be fresh and valued even if first produced today
  - The most widely acclaimed public domain exemplars are already online

- Current, directly applicable knowledge
  - Much (though not all) public domain informative literature now outdated (especially if one only looks pre-1923)

- Entertainment
  - Important, but hard to get grants for…

- History and documentation
  - Vast majority of interesting public domain text
  - Includes literature, art, essays, science, etc., that are valuable primarily in a historical context
Serials: A treasure trove of history and documentation

• Newspapers: First reports of events
  – Primary sources for history; essential for local history

• Magazines: Literature, essays, debates as they first appear
  – And often the only place that they appear
  – Lots of short-form work that didn’t make it into books

• Scholarly journals: The record of research
  – The hypotheses, the experiments, the data, the debates

• Specialty publications: Insight into communities
  – Trade journals, local and special interest groups
What’s in the public domain?

• Anything copyrighted before 1923
• Anything that’s specifically dedicated
  – E.g. fed. gov. docs in US; small amounts of private stuff
• Anything that didn’t “maintain” copyright as was once required:
  – Copyrights before 1964 that were not renewed (most weren’t)
    » But many of most significant books were renewed
    » (note also some may contain separately renewed material)
  – Publications before 1989 w/o copyright notice
    » (Inadvertent omissions after 1977 sometimes fixable)
  – But: Many foreign works were retroactively exempted from maintenance requirements in 1996
    » Key requirement: First foreign publication needs to be more than 30 days before first US publication to get exemption
Copyright detective work

- Popular Lord Peter Wimsey mystery in US public domain since 1951, but not online until 2005

- Needed to:
  - Search renewal records
  - Research publication history
  - Find a first edition to transcribe
  - Meticulously record what we did to clear the book

- Expensive to scale up
  - (see work by Denise Troll Covey at CMU)
Copyright renewals index

For periodicals, you may find a list of first copyright renewals for periodicals useful.

Renewal records, by year

Here is what is available online so far:

1950 - 1977 (renews copyrights from 1922 - 1950)

These years are not yet in an official database; however, volunteers have scanned renewal page images for many book-related categories. See also the transcriptions and indexes above. Books may be covered under copyrights in related categories. (For instance, if a portion of a book first appeared in a magazine, it may still be copyrighted under the magazine's copyright even if the book's copyright was not renewed.)

The page images can be found here:

- 1950 (everything except music)
- 1951 (books, periodicals, contributions to periodicals, drama, artwork)
- 1952 (books, periodicals, contributions to periodicals, drama, artwork)
- 1953 (books, periodicals, contributions to periodicals)
- 1954 (books, periodicals, contributions to periodicals)
- 1955 (books, periodicals, contributions to periodicals)
- 1956 (books, periodicals, contributions to periodicals)
Copyright renewal volumes

Renewals for Periodicals

The following page images are available (all GIFs):

- Part 1 (January - June):
  - Page 199 (A - Atlantic Monthly)
  - Page 200 (Atlantic Reporter - Chain Store Age)
  - Page 201 (Child Life - Fortune (start))
  - Page 202 (Fortune (cont.) - Mademoiselle)
  - Page 203 (Medical Clinics of North America - New York Herald Tribune (start))
  - Page 204 (New York Herald Tribune (cont.) - New York Times (start))
  - Page 205 (New York Times (cont.) - New Yorker (start))
  - Page 206 (New Yorker (cont.) - Pete Rice Magazine (start))
  - Page 207 (Pete Rice Magazine (cont.) - Saturday Review of Literature (start))
  - Page 208 (Saturday Review of Literature (cont.) - Street & Smith's Complete Stories (start))
  - Page 209 (Street & Smith's Complete Stories (cont.) - Street & Smith's Wild West Weekly (start))
  - Page 210 (Street & Smith's Wild West Weekly (cont.) - Vanity Fair)
  - Page 211 (Vogue - Western Winners)

- Part 2 (July - December):
ASTOUNDING STORIES. © Conde Nast Publications, Inc. (PCW)

3Apr62; v.14, no. 6, Feb35. © 16Jan35; B248817.
3Apr62; 12Feb62; R290834.
3Apr62; v.15, no. 1, Mar35. © 20Feb35; B253418.
3Apr62; 8May62; R294941.
ises,
3Apr35. © 20Mar35; B255773.
3Apr62; 8May62; R294971.
89877.
3Apr35. © 17Apr35; B259347.
3Apr62; 8May62; R294999.
52450.
4, Jun35. © 15May35; B261870.
54759. 4Jun62; R296180.
58231.
THE ATLANTIC MONTHLY. © Atlantic Monthly Co. (PCW)

3Apr62; v.154, no. 1, Jul34. © 13Jun34; B228379.
Inventory of serials

- Astonishing Stories: issues renewed from February 1940 (v. 1 no. 1); see 1967
- Astounding Science Fiction: see Astonishing Stories
- Astounding Stories: issues renewed from October 1933 (v. 12 no. 2); see 1960 Jul-Dec
- Athlete: issues renewed from August 1939 (v. 1 no. 1); see 1966
- The Atlantic Monthly: issues renewed from July 1934 (v. 154 no. 1); see 1962 Jan-Jun
- Atlantic Reporter: issues prior to 1923 renewed; see 1950 Jan-Jun
- Audubon Magazine: issues renewed from Sep-Oct 1947 (v. 49 no. 5); see 1975
- Authors' League Bulletin: issues renewed from June 1942 (v. 1 no. 3 -- new series? ALBs can be found back to 1913); see 1970
- The Avenger: issues renewed from September 1939 (v. 1 no. 1); see 1966
- Awful Oscar: "consult copyright office"; see 1977 Jan-Jun; first cataloged renewal is for June 1949 (v. 1 no. 11) recorded in 1977 Jul-Dec
- Baby Talk: issues renewed from Feb. 1938 (v. 3 no. 2); see 1965
- Bacteriological Reviews: issues renewed from June 1940 (v. 4 no. 2); see 1968
- Baptist Leader: "consult copyright office"; see 1973; first cataloged renewal is for January 1946 (v. 7 no. 10) recorded in 1977 Jul-Dec
- Barron's: issues renewed from January 6, 1941 (v. 21 no. 1); see 1968
- Barron's National Business and Financial Weekly: see Barron's
- Batman: issues renewed from Spring 1940 (v. 1 no. 1); see 1967; see also Detective Comics
- Battle Bird: see Battle Birds
- Battle Birds: issues renewed from February 1940 (v. 1 no. 1); see 1967
- Beauty Fair: issues renewed from Dec-Jan. 1947 [sic] (v. 2 no. 5); see 1976 Jan-Jun
- Beginners Teacher's Quarterly: issues renewed from October 1, 1941 (v. 28 no. 1); see
Very few serials completely copyrighted after 1922

- WorldCat lists over 200,000 serials with significant dates between 1923 and 1950
  - May undercount based on journals not starting or ending in that range
  - May overcount based on duplicate entries for some serials
- We found only ~1300 serials that renewed any issue copyrights during that time
- Most significant serials publishing then did not renew all (or any) of their issues
  - Weak correspondence between extent of renewal and significance to researchers
Newspapers

- No daily newspaper outside New York renewed issues dating before the end of World War II
- Only a few dailies from 1923-1950 renewed at all
- Earliest renewed issues for some major newspapers:
  - New York Herald/Tribune: before 1923
  - New York Times (daily): 1928
  - Wall Street Journal: 1941
  - Christian Science Monitor: 1945
  - Chicago Tribune: 1946
  - Washington Post: 1951
  - Los Angeles Times: 1958
  - Boston Globe, Philadelphia Inquirer, many more: No issue renewals
Scholarly journals

Using JSTOR (as of March 2006) as representative sample of significant journals:

– 1923-1950 journals in JSTOR: 298
– Number that renewed any issues: 49
– Number renewing first issue in period: 7
Magazines

- Some major magazines renewed from the start (or before 1923)
  - E.g. The New Yorker, National Geographic, Sat Eve Post
- Many other majors did not start renewing right away
  - E.g. Time, The New Republic, Scientific American
- Many others didn’t renew at all
  - E.g. The Economist
- Comics, pulp fiction often renewed aggressively
What’s published in the US?

THE ECONOMIST
ESTABLISHED 1843

No. 5342 JANUARY 12, 1946

No Assembly - - - 41 Remnants of a Race - - - 45
Matter of Men - - - 42 Opportunity for Light Metals - 60
Future Food Policy - - - 44 PAYE Simplified - - - 61

OF THE WEEK
- - - - 46 THE WORLD OVERSEAS - - - - 56
The Fourth Republic; Hesitations in Persia; India’s Labour Outlook; Canadian Export Policy.

THE BUSINESS WORLD
BUSINESS NOTES - - - - 62

Entered as Second Class Mail Matter at the New York, N.Y., Post Office

John Mark Ockerbloom UNIVERSITY OF PENNSYLVANIA LIBRARY Apr. 11, 2006
Caveats and qualifiers

• Serials may contain separately copyrighted (and separately renewed) items
  – Text: Contributions to periodicals (renewal scans and transcriptions are available online)
    » Inventory of these for serials would be useful too
  – Images: Renewal scans not yet online, but there aren’t that many
    » Low-hanging fruit for a scanning/transcription project!

• Some possible mitigating factors:
  – Section 201(c) gives serial copyright holders presumptive rights to reprint contributions in original context
  – Orphan works provisions might also make it easier to clear contributions

• Double-check anything you intend to digitize
  – Don’t rely on me (or other non-lawyers) for legal advice
  – It’s possible we may have missed renewals
Who will provide libraries of historic serials online?

• Google, Yahoo, Microsoft?
  – Serial content being ingested along with monograph content
  – Thus far not being treated specially or systematically

• Commercial aggregators?
  – Already broad coverage, but still limited; limited access; no inherent monopoly on public domain content

• Let central consortia do it?
  – JSTOR et al can’t do everything; and limited access may inhibit reuse, repurposing, remixing

• Libraries do it ourselves?
  – We have the content, constituency, mission, know-how
  – Scanner tech, storage, and sharing all getting cheaper
Sharing knowledge about copyrights

- Renewal scans were contributed by many, inventory created by individual
  - Just using flat files served by Apache
  - Thanks: Greg Weeks, Juliet Sutherland & Distributed Proofreaders, CMU, Penn, several public libraries

- Many may need to do copyright research
  - Searchable databases useful for quick lookups
  - Registries useful for pooling information

- Rights clearance info could also be machine-readable
  - See e.g. rights expression work by Karen Coyle, CDL
  - Needs to avoid inhibiting expression, contributors
Building a robust knowledge base for copyright information

• **Usability**: Should be easy for humans and programs to comprehend as needed
  – Main machine-processing requirement: Lookup

• **Inclusiveness**: Make it easy for researchers to contribute any relevant information
  – What work (journal, issues…) is being referred to?
  – What is asserted about its copyright?
  – What are the facts that support this assertion?
  – Who is making the assertion? (And perhaps: when?)

• **Reliability**: Support degrees of certainty and authority, audit (history) trail
Where and how to construct this knowledge base?

• Stick with individually curated flat files?
  – Certainly easy to store, trade; hard to scale

• Wiki base with structured data and review?
  – We built one not long ago: Fred (format registry demonstration that was prototype for GDFR)
    » http://tom.library.upenn.edu/fred/ for reference

• Add-on to existing catalog/registry?
  – E.g. Registry of Digital Masters on OCLC

• Should extended scope be considered?
  – Journals: Info on digitizations as well as copyright?
  – Copyrights: info on non-journal copyrights as well?

Who will use and support it, and what do they need?
To sum up

• We have many valuable, relatively recent serials that are fair game to digitize and share
  – An opportunity largely untapped to date (at least post-1922)
• We now have, or can get, the information we need to determine their status and share them digitally
• DLF and its member institutions are particularly suited to leading in this area
  – Due to our collections, the research we support, and our expertise coordination
• It would be useful, and feasible, to coordinate our efforts and our information gathering
  – E.g. by contributing to appropriate information registries
• We should begin discussions on how best to move forward
So, let’s discuss

• Initial inventory and copyright renewal records at
  http://onlinebooks.library.upenn.edu/cce/

• You can contact me at
  – Ockerblo@pobox.upenn.edu

• What information could you use? What could you contribute?
• Let’s start the conversation