Automated Rights Determination to Unlock Public Domain Treasures in Library Collections

Stephen Abrams  
Head of Digital Preservation

Suzanne Wones  
Associate University Librarian for Discovery and Access

Kyle K. Courtney  
Office of Scholarly Communication Copyright Advisor

Mingtao Zhao  
Systems Analyst and Application Developer

Harvard Library

Wednesday  
December 8, 2021
“The goal is to promote and encourage use and reuse of the collections by anyone for any purpose”

The problem

• More than 16 million bibliographic items – 7 million already digital or digitized
• But only 90,000 openly shared via DPLA
• Over 3 million created/published more than 130 years ago and likely in the public domain
• How can we know whether they are or not?
• We can manually review about 28,000/year, so it would take us only a mere 570 years to clear everything
The solution

• Automated rights determination for at-scale review based on pertinent catalog metadata
  • Ability to “dial-in” acceptable level of risk tolerance
  • Benchmark quality no worse than manual review

• Standardized rights statements
  https://rightsstatements.org/

• More information available at
  https://wiki.harvard.edu/confluence/display/digitalpreservation/Standardized+Rights+Statements
Algorithmic decision tree

• 18 decision points leading to 19 instances of 4 copyright statuses
  • No known copyright
  • No copyright – US
  • Copyright undetermined
  • In copyright
Algorithmic decision criteria

- Publication status
- Creation/publication date
- Corporate/individual authorship
- Death date
- Birth date

- Explicit policy of tolerating greater potential of false negatives (excluding from PD when it really is) to avoid any possibility of false positives (including in PD when it really isn't)
**Proof-of-concept**

- 60,000 item test set (drawn from 16 million MARC records)
- 130-year publication threshold for “obvious” public domain status

<table>
<thead>
<tr>
<th>Status</th>
<th>Published</th>
<th>Manuscript</th>
<th>Unknown</th>
<th>Total</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Known Copyright</td>
<td>9,366</td>
<td>–</td>
<td>–</td>
<td>9,366</td>
<td>15.6%</td>
</tr>
<tr>
<td>No Copyright – US</td>
<td>2,991</td>
<td>17</td>
<td>–</td>
<td>3,008</td>
<td>5.0%</td>
</tr>
<tr>
<td>Copyright Undetermined</td>
<td>271</td>
<td>2</td>
<td>3,847</td>
<td>4,120</td>
<td>6.9%</td>
</tr>
<tr>
<td>In Copyright</td>
<td>43,138</td>
<td>277</td>
<td>–</td>
<td>43,415</td>
<td>72.4%</td>
</tr>
<tr>
<td>Excluded licensed e-resources</td>
<td>91</td>
<td>–</td>
<td>–</td>
<td>91</td>
<td>0.2%</td>
</tr>
</tbody>
</table>

**Total** 55,857 296 3,847 60,000
Proof-of-concept

• Using a more conservative 140-year threshold did not materially affect the results
  • 3 items fell into In Copyright
  • 530 moved from No Known Copyright to No Copyright – US
“What could possibly go wrong?”

- Small sample size – representative, but only 0.375%
- Reliance on metadata never intended for this purpose
- Missing, incomplete, non-uniform, incorrect metadata

<table>
<thead>
<tr>
<th>Year</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1926</td>
<td>Tan’gi 4300 [1967]</td>
</tr>
<tr>
<td>1926 January</td>
<td>Nahāsē 2009 ’Ā. Me [August 2017]</td>
</tr>
<tr>
<td>1926 January 7</td>
<td>1377 [1998 or 1999]</td>
</tr>
<tr>
<td>&lt;1926&gt;</td>
<td>1390- [2011-487 i.e. 1727]</td>
</tr>
<tr>
<td>(1926)</td>
<td>---</td>
</tr>
<tr>
<td>1926</td>
<td>1926?</td>
</tr>
<tr>
<td>1926 (?)</td>
<td>1---</td>
</tr>
<tr>
<td>circa 1926</td>
<td>18---</td>
</tr>
<tr>
<td>ca. 1926</td>
<td>185-</td>
</tr>
<tr>
<td>c1926</td>
<td>uuuu</td>
</tr>
<tr>
<td>approximately 1926</td>
<td>1uuu</td>
</tr>
<tr>
<td>approx. 1926</td>
<td>18uu</td>
</tr>
<tr>
<td>flourished 1920</td>
<td>185u</td>
</tr>
<tr>
<td>fl. 1920</td>
<td>----</td>
</tr>
<tr>
<td>active 1850</td>
<td>1920 or 1921</td>
</tr>
<tr>
<td>active ca. 1850</td>
<td>1920 or 21</td>
</tr>
<tr>
<td>born 1923</td>
<td>1920/21</td>
</tr>
<tr>
<td>b. 1923</td>
<td>1893 to 1994</td>
</tr>
<tr>
<td>died 1982</td>
<td>between 1893 and 1994</td>
</tr>
<tr>
<td>d. 1982</td>
<td></td>
</tr>
</tbody>
</table>
Risk mitigation

• Extensive algorithmic rules for metadata normalization and (conservative) error interpretation
• Easily parameterizable threshold dates
Risk mitigation

• Legal nuance, e.g.,
  • Twin Books case problem
  • Subsequent publication of manuscripts
  • Possible copyright renewal
  • “Inserts” in materials
  • Geoblocking
  • Etc.
Next steps

• Application to all 16 million bibliographic items
• Release of a rights-cleared public domain corpus
• Implement addition of SRS at the time of cataloging
• Extend the algorithm for application to all 25 million cataloged items, e.g., non-MARC, visual, archival, etc.
• Investigate integration with external authority DBs (e.g., VIAF) for better author birth/death information
• Investigate automated identification of in-volume copyright notices
Summary

• Validation of automated approach to rights determination at scale
  • Technically possible
  • Reliable results
• Rights determinations represent (as they always have) our best due diligence based on available information
• Potential for significant opening-up of deep collections for public use

https://wiki.harvard.edu/confluence/display/digitalpreservation/Standardized+Rights+Statements
Thanks to our colleagues!

• Wendy Gogel  Manager of Digital Content and Projects
• Peter Hirtle  Independent scholar
• Jonathan Hulbert  Office of General Counsel
• Vanessa Venti  Library Analyst for Stewardship of Digital Assets
• Robin Wendler  Metadata Analyst
Questions?

Stephen Abrams  stephen_abrams@harvard.edu
Suzanne Wones  suzanne_wones@harvard.edu
Kyle K. Courtney  kyle_courtney@harvard.edu
Mingtao Zhao  mingtao_zhao@harvard.edu

https://wiki.harvard.edu/confluence/display/digitalpreservation/Standardized+Rights+Statements