Moving Beyond the Prologue: Exploring Data-centered Approaches to Prioritizing Serials Preservation

Charleston Conference 2017

Amy Wood
Head of Technical Services, Center for Research Libraries

Stephanie Orphan
Director of Publisher Relations, Portico

Craig Van Dyck
Executive Director, CLOCKSS
Collaboration to Preserve “Long Tail” Publishers’ Publications

Portico and CLOCKSS are the leading community-based digital preservation services for scholarly materials. Both have global footprints and have been operating for 10-15 years. They work with academic libraries and scholarly publishers to ensure future access to scholarly materials.

The Center for Research Libraries is an international consortium of university, college, and independent research libraries. Founded in 1949, CRL supports original research and inspired teaching in the humanities, sciences, and social sciences by preserving and making available to scholars a wealth of rare and uncommon primary source materials from all world regions.
Agenda

• CRL’s data centered approach to identifying and preserving the critical corpus of humanities & social sciences print journals

• Challenges in chasing the long-tail of E-publishers

• Ways in which CLOCKSS, Portico and CRL propose to work together

• Role of other related resources
CRL Print Serials Preservation Timeline

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Acquiring, Storing & Reformatting**
- Deposits, Acquisition
- Programs, Sharing
- Resources

**Preserving American Print Resources**

**Strengthen the Network of Involved Institutions**
- Support Informed Decision Making

**Data driven Approach**

**Critical Corpus**

**GPO Matrix**

**PAPR**

**PAN Forum**

**@RISK**

Midwest Inter-Library Corporation (MILC)
Problem:
Small percentage of overall corpus of print serials has been archived

Obstacle to serial archiving success:
Absence of consensus on the appropriate set of materials to be preserved

Solution:
Define the critical corpus by leveraging the historical knowledge and expertise invested by American research libraries over the past two centuries
Critical Corpus Project

Goal:
To define the costs and requirements for preserving the “universe” of print Humanities and Social Sciences serials.

Process:
• To develop and cost out a methodology and strategy to identify the “critical corpus” of print humanities and social sciences journal literature.
• To develop a significantly large list of titles to lay the groundwork for review and curation of a final list.
Requested full MARC records from Columbia, Cornell, NYPL & Princeton

Added previously processed records from 13 BTAA libraries

Downloaded records from PAPR

2.5 M Records
462,850 unique HSS titles were identified

- B - Philosophy, Psychology, Religion: 26,585
- C - Auxiliary Science of History: 5,386
- D - World History and History of Europe, Asia,...: 51,636
- E - History of the Americas: 7,157
- F - History of the Americas: 17,130
- G - Geography, Anthropology, Recreation: 18,306
- H - Social Sciences: 212,096
- J - Political Science: 25,298
- M - Music and Books on Music: 9,456
- N - Fine Arts: 20,334
- P - Language and Literature: 49,466
- Additional titles presumed SSH: 19,999
Digital Archive Status of Unique Titles

Critical Corpus HSS titles

<2% in CLOCKSS or Portico

Print Only

Digital
Lack of Digital Puts Print Preservation at Risk

• Serials preservation efforts are only sustainable when coupled with access.
• Serials are more likely to be used and valued if there is digital copy.
• Preserving digital copy in a trusted repository—CLOCKSS & Portico—will prevent loss of critical resources and ensure continued access.
• Need to plan and prioritize acquiring and archiving digital.
CLOCKSS/Portico and the Long Tail

• As community-based preservation services, CLOCKSS & Portico are committed to working with publishers across the spectrum of scholarly publishing

• current focus = “long tail” (10 or fewer titles)

• Content from long tail more at risk, preservation imperative, but there are challenges . . .
  • More long tail publishers than can be managed
  • Many are obscure, often we don’t know they exist
  • Costly to work with
Universe of Small Publishers

# of publishers with 10 or fewer (online) titles

- CRL Critical Corpus: 8,970
- CrossRef: 9,306
- Ulrichs: 6,662

Small pubs preserving

- with CLOCKSS: 182
- with Portico: 243

Publishers preserved in CLOCKSS/Portico by size

- 1 e-journal title: 37%
- 2 to 10 e-journal titles: 30%
- More than 10 e-journal titles: 33%
Evolving Strategies

• Current strategies for publisher preservation include
  • Targeting pubs using known platforms
  • Expansion of existing relationships (growth through 3\textsuperscript{rd}-party publishers)
  • Removing technical barriers
  • Seek partners for collaboration

• CRL Critical Corpus Project provides model for data-centered prioritization, opportunity for collaboration
Seeds of a Data-Centered Collaboration

- CRL, CLOCKSS, and Portico investigating ways to leverage & expand on the existing data
  - Analyze HSS e-serial titles not archived in CLOCKSS, Portico
  - Explore hosting a scholars’ evaluation of the critical corpus methodology and plan, and support for expanding/refining the list
  - Investigate and execute ways to engage community to discuss how to refine/correct the list (e.g. eliminate “false positives”)
  - Further refine/prioritize list of titles for preservation:
    - Important titles—most widely held (based on CRL data)
    - Endangered titles—least widely held (based on CRL data)
    - Expedient titles—publishers with multiple titles
Collaborative Approaches

• **Goals of collaboration** between CRL, CLOCKSS and Portico
  • develop a **prioritized list** of titles to pursue
  • create a **plan** to pursue the top-priority titles

• Input from the library community is crucial—share your thoughts/ideas, share your lists

• Opportunities to expand our research by using other tools, such as Keepers Registry
Preserving the “Long Tail”

Publishers must qualify as reputable, and the content must be considered preservation-worthy:

- Trusted 3rd-parties help to evaluate publishers
- Beall’s List
- Cabell’s white and black lists
- Librarians’ input
The Keepers Registry

• The Keepers Registry ([https://thekeepers.org](https://thekeepers.org)) reports on what journals are in which archives

• Developed by EDINA - a center for digital expertise and online service delivery at the University of Edinburgh - and the ISSN International Center

• Hosted by EDINA, and funded by Jisc

• Archiving information for 36,000 titles, from 13 archiving organizations
The Keepers

- Archaeology Data Service
- British Library
- Cariniana
- CLOCKSS
- Global LOCKSS Network
- HathiTrust
- Library of Congress
- National Library of the Netherlands
- NSL, Chinese Academy of Sciences
- PKP PLN
- Portico
- Swiss National Library
- Scholars Portal
Prioritizing the long tail

- Keepers Registry data can be helpful in identifying titles to prioritize for archiving
- Libraries and others can use the Title List Comparison to see the archival status of a list of titles
- Prototype statistics portal provides information on the archival status of a number of journal datasets
Title List Comparison

1. Create a list of titles

2. Submit list to Keepers Registry

3. Review statistics and analyze results

Key summary statistics:

% of unique titles which are being archived by three or more Keepers: 56.47%.

% of unique titles with e-journal records for which no archiving activity could be identified and which therefore may be at risk of loss: 26.95%.
Statistics Portal Prototype

- Could help with prioritizing the “long tail”
- Combines the data held in the Keepers Registry with holdings of UK libraries, and provides information about the archival status of the titles
- [http://labs.edina.ac.uk/keeperstats/#/](http://labs.edina.ac.uk/keeperstats/#/)
The REF, or Research Excellence Framework, is a system to assess the quality of research and outcomes in UK Higher Education institutions.

Submissions to the REF include up to 4 research outputs for each staff member.

Journals in which these research outputs were published were compared against the data in the Keepers Registry.
# Statistics Portal – REF2014

<table>
<thead>
<tr>
<th>Title</th>
<th>No. Titles</th>
<th>1+</th>
<th>3+</th>
<th>Ingest Ratio</th>
<th>KeepSafe Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classics</td>
<td>371</td>
<td>250</td>
<td>67</td>
<td>67.39</td>
<td>18.06</td>
</tr>
<tr>
<td>Law</td>
<td>3169</td>
<td>2174</td>
<td>1635</td>
<td>68.60</td>
<td>51.59</td>
</tr>
<tr>
<td>Art and Design: History, Practice and Theory</td>
<td>1413</td>
<td>1075</td>
<td>810</td>
<td>76.08</td>
<td>57.32</td>
</tr>
<tr>
<td>Modern Languages and Linguistics</td>
<td>2116</td>
<td>1662</td>
<td>1146</td>
<td>78.54</td>
<td>54.16</td>
</tr>
<tr>
<td>English Language and Literature</td>
<td>2283</td>
<td>1808</td>
<td>1126</td>
<td>79.19</td>
<td>49.32</td>
</tr>
<tr>
<td>Music, Drama, Dance and Performing Arts</td>
<td>1113</td>
<td>886</td>
<td>588</td>
<td>79.60</td>
<td>52.83</td>
</tr>
<tr>
<td>Theology and Religious Studies</td>
<td>529</td>
<td>423</td>
<td>242</td>
<td>79.96</td>
<td>45.75</td>
</tr>
<tr>
<td>Public Health, Health Services and Primary Care</td>
<td>4343</td>
<td>3625</td>
<td>2068</td>
<td>83.47</td>
<td>47.62</td>
</tr>
<tr>
<td>Area Studies</td>
<td>859</td>
<td>736</td>
<td>550</td>
<td>85.68</td>
<td>64.03</td>
</tr>
<tr>
<td>History</td>
<td>2606</td>
<td>2235</td>
<td>1481</td>
<td>85.76</td>
<td>56.83</td>
</tr>
</tbody>
</table>
Statistics Portal – OpenURL Router

• The OpenURL Router is a service used by around 100 UK universities to direct user requests to an appropriate local resolver

• Requests are made by researchers and students, and provide a sample snapshot of the journal titles that are being used
# Open URL service data for full year of 2015

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. Occurrences</td>
<td>4085133</td>
</tr>
<tr>
<td>No. ISSN-L</td>
<td>51426</td>
</tr>
<tr>
<td>1+</td>
<td>21461</td>
</tr>
<tr>
<td>3+</td>
<td>12126</td>
</tr>
<tr>
<td>Ingest Ratio</td>
<td>41.73</td>
</tr>
<tr>
<td>KeepSafe Ratio</td>
<td>23.58</td>
</tr>
<tr>
<td>Archiving Status Unknown</td>
<td>58.27</td>
</tr>
</tbody>
</table>
Take Away

Goal:
Preserve the content of Long Tail Publishers

Important Elements:
• Obtain library community input
• Prioritize
• Develop a plan
• Collaborate to succeed
Contacts

Amy Wood, Center for Research Libraries
email: wood@crl.edu

Stephanie Orphan, Portico
email: Stephanie.Orphan@ithaka.org

Craig Van Dyck, CLOCKSS
email: cvandyck@clocks.org