Preserving America's Print Resources II:

a North American Summit
Welcome and Overview:
Where We Are and What Has Changed Since the 2003 PAPR Conference

Bernard Reilly
President, Center for Research Libraries
Part 1: The Data In Hand and What the Data Tell Us
The Data In Hand

Amy Wood (Director, Technical Services, CRL)
Constance Malpas (Research Scientist, OCLC)
Complementary Perspectives

CRL
North American perspective
Program support & analysis
PAPR registry

Infrastructure purpose-built for item-level aggregation

Challenge:
Sustain specialized infrastructure
Encourage community-wide transparency of conditions & risk

OCLC
‘System-wide’ perspective
Aggregation, processing @ scale
Utility service provider

Infrastructure designed for title-level cooperative cataloging

Challenge:
Support emerging service needs within existing cooperative infrastructure

Better together
Coverage of Shared Print Journals

**PAPR**

- 38 registered programs
- 14 active programs, 110 libraries
- 50.2K unique titles
- 89K records & holdings

74 facilities described
- Portico
- CLOCKSS

**64 registered shared print symbols**

- 40 active contributors (journals)
- 54K shared print serial titles
- 71K shared print journal holdings
- 1.4M shared print book titles

Total library holdings on shared print journal titles: **8.4M**

measure of potential impact

December 2014
Sources of Shared Print Journal Holdings in WorldCat
December 2014

Largest individual repositories:
- SUNY Buffalo (28K titles)
- FLARE (22K)
- CIC Shared Repository (24K)
- Stanford (2.1K)
- CRL JSTOR (1.8K)

Largest programs: % of Aggregate Archive
- Empire Shared Collections: 53%
- Scholars’ Trust: 41%
- WEST: 19%
- Maine Shared Collections: 11%
- CIC Shared Print: 4%

N = 53,610 titles; 40 active repositories registered in WorldCat
CRL’s Analysis Included

PAPR database - ~50,200 unique titles (~89,000 records)

ProQuest’s Ulrichs database - ~436,130 unique titles

HathiTrust - ~292,480 unique titles

CRL’s catalog - ~52,212 unique titles

Portico - ~16,247 unique titles with preserved status

JSTOR - ~3,300 unique titles
Data Points Considered in CRL Analysis

- Title Identifiers – ISSN, OCLC#, Database Record IDs
- LC Classification
- Beginning and end publication dates
- Language
- Country of Publication
- Holdings
- Print Archiving Program
- Institution / Library

- 22 supporting or secondary values
- Volume level data generated from holdings
- Volume level data generated from publication dates
Over past 8 years, the number of serial titles in WorldCat has tripled.
yet the proportion of serials has remained relatively stable
Overall decline in holdings: title is driven in part by increased scope and ‘internationalization’ of titles in WorldCat – loading of national bibliographies, etc.

... not evidence of large-scale inventory reduction
Barriers To Success

- No ISSN or Incorrect ISSN
- Incomplete Bib Records & Holdings
- Local & Legacy Data Practices
- Expecting One Size Fits All Analysis
Fortune favors the prepared (?)

The presence of standard identifiers facilitates cooperative action; **titles with ISSN** may have greater archiving potential.

- 72% of titles have ISSN.
- 28% of serial titles are associated with any ISSN data.

Cf. 82% of PAPR titles have ISSN.

N = 5.5M titles
Challenges

• Serials a ‘baggy monster’ of publication types and formats
• Holdings maintenance managed in multiple, overlapping environments – ILS, KB, union catalogs
• Aggregation of title-level holdings lacks detailed information on publication history, local holdings etc.
• Clustering and deduplication of titles complicated by sparse data
Critical Obstacles

• Redundant registration processes, infrastructure
• Reluctance to disclose group-scale commitments beyond participating members
• Descriptive practices vary: disclosure at title level vs. item level; local shared print holdings on multiple symbols etc.
• Cooperative stewardship operationalized in legacy institutional workflows
What the Data Tell Us

Amy Wood - Director Technical Services, CRL
Constance Malpas - Research Scientist, OCLC
<table>
<thead>
<tr>
<th>LC Class top level</th>
<th>OVERLAP by ISSN</th>
<th>PAPR only (ISSN)</th>
<th>PAPR (no ISSN)</th>
<th>ULRICHS only (ISSN)</th>
<th>ULRICHS (No ISSN)</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>76</td>
<td>72</td>
<td>69</td>
<td>531</td>
<td>180</td>
<td>852</td>
</tr>
<tr>
<td>E</td>
<td>143</td>
<td>140</td>
<td>86</td>
<td>837</td>
<td>159</td>
<td>1,221</td>
</tr>
<tr>
<td>U</td>
<td>80</td>
<td>316</td>
<td>89</td>
<td>1,117</td>
<td>273</td>
<td>1,795</td>
</tr>
<tr>
<td>F</td>
<td>315</td>
<td>300</td>
<td>232</td>
<td>2,827</td>
<td>1,300</td>
<td>4,658</td>
</tr>
<tr>
<td>C</td>
<td>64</td>
<td>207</td>
<td>64</td>
<td>3,438</td>
<td>1,220</td>
<td>4,929</td>
</tr>
<tr>
<td>M</td>
<td>185</td>
<td>127</td>
<td>48</td>
<td>3,681</td>
<td>1,779</td>
<td>5,634</td>
</tr>
<tr>
<td>N</td>
<td>313</td>
<td>481</td>
<td>129</td>
<td>8,878</td>
<td>3,212</td>
<td>12,699</td>
</tr>
<tr>
<td>D</td>
<td>699</td>
<td>565</td>
<td>189</td>
<td>9,901</td>
<td>2,387</td>
<td>13,041</td>
</tr>
<tr>
<td>J</td>
<td>339</td>
<td>320</td>
<td>216</td>
<td>9,961</td>
<td>5,341</td>
<td>15,837</td>
</tr>
<tr>
<td>Z</td>
<td>633</td>
<td>650</td>
<td>386</td>
<td>10,896</td>
<td>4,544</td>
<td>16,475</td>
</tr>
<tr>
<td>L</td>
<td>596</td>
<td>413</td>
<td>199</td>
<td>11,861</td>
<td>4,552</td>
<td>17,025</td>
</tr>
<tr>
<td>A</td>
<td>437</td>
<td>515</td>
<td>444</td>
<td>10,596</td>
<td>6,662</td>
<td>18,216</td>
</tr>
<tr>
<td>B</td>
<td>748</td>
<td>471</td>
<td>237</td>
<td>13,029</td>
<td>4,696</td>
<td>18,433</td>
</tr>
<tr>
<td>K</td>
<td>1,646</td>
<td>1,268</td>
<td>1,449</td>
<td>11,449</td>
<td>3,900</td>
<td>18,066</td>
</tr>
<tr>
<td>P</td>
<td>1,235</td>
<td>809</td>
<td>164</td>
<td>15,147</td>
<td>5,345</td>
<td>21,464</td>
</tr>
<tr>
<td>S</td>
<td>1,366</td>
<td>1,409</td>
<td>1,185</td>
<td>15,478</td>
<td>5,477</td>
<td>23,549</td>
</tr>
<tr>
<td>G</td>
<td>743</td>
<td>644</td>
<td>260</td>
<td>18,973</td>
<td>9,631</td>
<td>29,507</td>
</tr>
<tr>
<td>R</td>
<td>2,461</td>
<td>1,526</td>
<td>205</td>
<td>32,365</td>
<td>6,011</td>
<td>40,106</td>
</tr>
<tr>
<td>Q</td>
<td>7,905</td>
<td>5,915</td>
<td>2,851</td>
<td>35,173</td>
<td>8,961</td>
<td>52,899</td>
</tr>
<tr>
<td>T</td>
<td>5,409</td>
<td>4,781</td>
<td>3,922</td>
<td>38,398</td>
<td>15,397</td>
<td>62,497</td>
</tr>
<tr>
<td>H</td>
<td>2,743</td>
<td>2,245</td>
<td>1,147</td>
<td>69,531</td>
<td>32,376</td>
<td>105,298</td>
</tr>
</tbody>
</table>
PAPR overlap with Ulrichs

Uneven Coverage Across LC Class

~485,000 titles
Serials
12.1M titles

Periodicals excluding microforms, e-formats, braille etc.

5.5M titles

A reasonable proxy for the corpus mostly likely to be ‘in scope’ for print journal archiving efforts?

Perspective
As of December 2014, \( \sim 1\% \) of periodical titles are associated with one or more print archiving programs.
PAPR & Ulrichs Inside the Universe of Worldcat Print Serials

5.5 million titles
Publications from the 19th-21st centuries account for >86% of serial titles in WorldCat

Cataloging ‘backlogs’ insufficient to account for decline since mid-1990s?
Distribution of Ulrichs & PAPR by First Year of Publication
Subject distribution is difficult to assess

N = 5.5M titles

- Lack LCCN: 11% (89% of titles have LCCN)
- Lack Conspectus heading: 72% (28% have Conspectus heading)
- Lack text string subject heading: 64% (36% have text string subject heading)

OCLC analysis focused here
Distribution of Number of Titles by LC Classification

ULRICHs

HathiTrust

PAPR

Legend:
- A
- B
- C
- D
- E
- F
- G
- H
- J
- K
- L
- M
- N
- P
- Q
- R
- S
- T
- U
- V
- Z

0 50,000 100,000 150,000 200,000 250,000 300,000 350,000 400,000 450,000 500,000
Subject Distribution of Serials* in WorldCat for which Conspectus is Assigned

N = 1.5M titles

Business & Economics, 258129
History & Auxiliary Sciences, 135260
Education, 83341
Library Science, Generalities & Reference, 144646
Engineering & Technology, 110618
Archived corpus is **not representative** of system-wide print journal collection
<table>
<thead>
<tr>
<th>Language</th>
<th>Ulrichs Percentage</th>
<th>HathiTrust Percentage</th>
<th>PAPR Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>60.1%</td>
<td>66.4%</td>
<td>78.5%</td>
</tr>
<tr>
<td>German</td>
<td>8.8%</td>
<td>6.2%</td>
<td>Russian 4.5%</td>
</tr>
<tr>
<td>French</td>
<td>5.6%</td>
<td>5.0%</td>
<td>German 4.4%</td>
</tr>
<tr>
<td>Spanish</td>
<td>3.9%</td>
<td>3.6%</td>
<td>French 3.6%</td>
</tr>
<tr>
<td>Dutch</td>
<td>3.1%</td>
<td>Russian 3.0%</td>
<td>Spanish 2.1%</td>
</tr>
<tr>
<td>Italian</td>
<td>2.6%</td>
<td>Chinese 2.5%</td>
<td>Japanese 1.0%</td>
</tr>
<tr>
<td>Chinese</td>
<td>2.5%</td>
<td>Japanese 2.2%</td>
<td>Italian 0.9%</td>
</tr>
<tr>
<td>Japanese</td>
<td>1.5%</td>
<td>Italian 1.6%</td>
<td>Chinese 0.7%</td>
</tr>
<tr>
<td>Undetermined</td>
<td>1.3%</td>
<td>Portuguese 0.9%</td>
<td>Polish 0.7%</td>
</tr>
<tr>
<td>Portuguese</td>
<td>1.2%</td>
<td>Dutch 0.8%</td>
<td>Portuguese 0.5%</td>
</tr>
<tr>
<td>Russian</td>
<td>1.2%</td>
<td>Polish 0.8%</td>
<td>Dutch 0.4%</td>
</tr>
<tr>
<td>Danish</td>
<td>1.2%</td>
<td>Czech 0.6%</td>
<td>Romanian 0.4%</td>
</tr>
<tr>
<td>Swedish</td>
<td>1.1%</td>
<td>Korean 0.5%</td>
<td>Czech 0.4%</td>
</tr>
<tr>
<td>Polish</td>
<td>0.8%</td>
<td>Indonesian 0.5%</td>
<td>Swedish 0.2%</td>
</tr>
<tr>
<td>Norwegian</td>
<td>0.6%</td>
<td>Swedish 0.5%</td>
<td>Bulgarian 0.2%</td>
</tr>
</tbody>
</table>
N = 53,610 titles

Language Distribution of Serial Titles in Shared Print Repositories
December 2014

Archiving efforts have favored English language content

- English: 84%
- German: 4%
- French: 4%
- Spanish: 2%
- Italian: 1%
- Other: 5%
N = 5.5M titles

Holdings Distribution for Serial* Titles in WorldCat
January 2015

- 88% sparsely held

Avg. WorldCat holdings per title = 5
Median WorldCat holdings per title = 1
Holdings Distribution for Shared Print Serial Titles
December 2014

N = 53,610 titles

Archived corpus mostly represents ‘core’ titles

44% widely held

Avg. WorldCat holdings per title = 156
Median WorldCat holdings per title = 81
N = 53,610 titles

Duplication of Serial Titles in Shared Print Repositories
December 2014

Limited duplication across individual repositories

77% held by single repository
Number of Copies of Titles Archived by LC Classification

- 17 copies or records
- 12 copies or records
- 11 copies or records
- 10 copies or records
- 9 copies or records
- 8 copies or records
- 6 copies or records
- 5 copies or records
Number of Volumes held and reported in PAPR compared to Number of Volumes, for the same titles, published in LC Class V
PAPR Comparison of Volumes Held and Volumes Published

- Held
- Published

Years:
Comparison of Number of Volumes held vs. Number of Volumes Published of 147,683 titles in HathiTrust Database
Cost/benefit tradeoffs vary by discipline - strategies for securing widely held literature will be different from strategies for preserving specialized literature.
Absence of ‘above-consortium’ coordination will produce unexpected strengths (and weaknesses) in aggregate archive.
Conclusions

- Group-scale coordination of archiving efforts has achieved **measurable progress** in short time.
- Benefits would be more widely shared if archive metadata were more **systematically aggregated**.
- Collective shared print archive represents, **fraction of system-wide collection**, concentrated in a few subject areas representing core holdings.
- Cooperative archiving of less widely held titles will require more **conscious coordination**.
Areas for collective action

• More attention to role of identifiers in aggregation, selection, analytics

• Clarification of disciplinary/topical responsibility to preservation of scholarly record Greater incentives to register shared print holdings in network level aggregations

• Agreement on ‘common denominator’ for assessing coverage, progress