Managing Collections in the Networked Environment: New Analytic Approaches

OCLC Research Webinar
9 September 2010

Constance Malpas, OCLC Research
Zack Lane, Columbia University
Helen Look, University of Michigan
Jacob Nadal, University of California, Los Angeles
Context

• Making library data “work harder”
• Decision support: where should limited institutional resources be directed?
• New skill sets, professional cohort emerging

• Highlight significant work at RLG partner institutions
• Identify shared research priorities, methodologies
• Staffing and infrastructure requirements, organizational development
Today’s Panelists

Zack Lane
ReCAP Coordinator
Columbia University
Zl2114@columbia.edu

Helen Look
Collection Analyst
University of Michigan
hlook@umich.edu

Jacob Nadal
Preservation Officer
UCLA
jnadal@library.ucla.edu
Circulation Data at Columbia University Libraries

Zack Lane
ReCAP Coordinator
Columbia University
Project: Look at System-wide Circ Data

- Several data categories
- Many ways to slice
- Tip of the iceberg
Lots of Amazing (Granular) Data

- Main categories: accessions, retrieval, delivery and circulation

ReCAP Data Center

ReCAP Data Center website collects, defines and analyzes data related to Columbia University Library collections in ReCAP. Data are collected quarterly. For more information on individual categories please see executive summary.

A current project is underway to make all data publicly accessible. Contact Zack Lane, ReCAP Coordinator, with rec

<table>
<thead>
<tr>
<th>Raw Data</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accession</td>
<td>Retrieval rate</td>
</tr>
<tr>
<td>Retrieval</td>
<td>Timing of Requests</td>
</tr>
<tr>
<td>Delivery</td>
<td>High use titles</td>
</tr>
<tr>
<td>Circulation</td>
<td>High volume requests</td>
</tr>
<tr>
<td>CUL Bib File Statistics</td>
<td>EDD trends</td>
</tr>
<tr>
<td></td>
<td>Publication Date (Monographs)</td>
</tr>
<tr>
<td></td>
<td>Customer Code Correspondence</td>
</tr>
</tbody>
</table>

2) CUL Accession Data: CUL retrieves weekly batch files of new accessions from ReCAP. Using batch file CUL staff update CLIO location and request button in the OPAC. Granular data about each item is available via Voyager.

- Raw accession data (txt): Compiled by CUL per month
- Categories: Barcode, Customer Code, Accession Date, Bib ID, Format, Publishing Date, Language, Title, Holdings ID, Call Number, Enumeration/Chronology, Item ID, CLIO Location, Year of Accession, Month of Accession and Day of Accession
Managing Collections in the Networked Environment: New Analytic Approaches

Multi-dimensional Analyses
Circulation Analysis Project: Spring 2010

• Identify bright and capable intern: Steve Zweibel!
• Locate data sets
• Understand data sets
• Working with Systems staff to improve data
• Reformatting Data
• Manipulating data with Excel 2003/2007 (Pivot tables)
• Presenting data with Power Point
• Rethink, rework and refine
Big Numbers: 18% decline in circ over 7 years

Total Charges by FY

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>FY03/04</td>
<td>570827</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FY04/05</td>
<td>567739</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FY05/06</td>
<td>527171</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FY06/07</td>
<td>505288</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FY07/08</td>
<td>474218</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FY08/09</td>
<td>469217</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FY09/10</td>
<td>467272</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
What’s to Gain?

• Data set is clean and compact
• Categories apply to both on-site and off-site collections
• Low barrier of entry for library school intern
• Everyone has access; no-one is looking
• Improves understanding of circ system, patron behavior, staff habits, data analysis tools and system-wide trends
• Staff somewhat surprised that circulation data analysis was system-wide not ReCAP-specific
Big Numbers: Revisited

Charges by Patron Group: Monthly

4 main patron groups: Grad Students, Faculty, Undergrads and Visitors
Zack’s Deep Thoughts

- Trend of total charges is downward due to use of e-resources
- Leveling off indicates that use of print copy is still strong and critical
- Faculty charges have increased due to more Grads serving as adjunct faculty (with OFF privileges)
- Grads peak a month before Undergrads because of course requirements (tilted towards written papers instead of tests)
What Else Did Steve Discover?

Renewals by Patron Group: Monthly

Patrons renew when they must renew
Patron Group Habits:

- Charge/Renewal ratios are consistent with staff perceptions

- Faculty are more likely to hold onto the books that they have *than charge* out new ones

- Undergraduates have little need to renew with extended loan period

- Grad students charge a lot of material and hold it for more than one term

- Visitors hold books longer since obtaining **OPAC renewal permission**
Onsite Vs. Offsite Charges

- **Onsite**
- **Offsite**

<table>
<thead>
<tr>
<th>Year</th>
<th>Offsite</th>
<th>Onsite</th>
</tr>
</thead>
<tbody>
<tr>
<td>FY03/04</td>
<td>20000</td>
<td>500000</td>
</tr>
<tr>
<td>FY04/05</td>
<td>15000</td>
<td>550000</td>
</tr>
<tr>
<td>FY05/06</td>
<td>12000</td>
<td>520000</td>
</tr>
<tr>
<td>FY06/07</td>
<td>20000</td>
<td>500000</td>
</tr>
<tr>
<td>FY07/08</td>
<td>18000</td>
<td>520000</td>
</tr>
<tr>
<td>FY08/09</td>
<td>17000</td>
<td>510000</td>
</tr>
<tr>
<td>FY09/10</td>
<td>16000</td>
<td>500000</td>
</tr>
</tbody>
</table>

Data indicates a decline in onsite charges from FY03/04 to FY09/10, with a corresponding increase in offsite charges.
Relevant to Other Issues

Access Services considering changes to Hold policy:

• Holds have limited duration
• Data indicates that most Holds expire
• Should Hold duration be extended?
• Should Holds (via OPAC) be eliminated?

Enter Circ data analysis...
Overall Holds are half that of Recalls

Online Holds are one-eighth that of Recalls
Moving Forward

• **Bring data to staff;** don’t expect staff to come to data
• **Learn** from the intern
• **Query, pressure and improve** data

• **Congratulations Steve Zweibel!** Part-time professional cataloging position at NYU and part-time reference at Hunter College Health Sciences Library
Post-digitization Use of
Print Collections

Helen Look

Collection Analyst
University of Michigan
Background

• Study of post-digitization use of print collections at the University of Michigan

• University of Michigan digitization efforts

• HathiTrust Digital Library - [http://www.hathitrust.org](http://www.hathitrust.org)
Access to Institutional Resources

- Harmonizing data from different sources
- Working with different staff to gather the data
- Consulting with internal and external colleagues
Methodology

• Top 500 accessed titles in HathiTrust Digital Library by the University of Michigan community in 2009

• Title-level online usage was compared to title-level print usage

• Print circulation for the sample was compiled for 2008, 2009 and total circulation history
Low Usage of the Print

2009 Circulation

- 98% (489) of the titles had zero circulation
- 2% (11) of the titles circulated
- 2009 circulation for the 11 titles was equal to or less than the 2008 circulation
Increased Discoverability of the Content

Historic Circulation

- 39% (193) of the titles had not circulated
- 61% (307) of the titles had circulated
- Hidden treasures
Subject Distribution

- Agriculture: 2%
- Auxiliary Sciences of History: 1%
- Bibliography, Library Science, General Info Resources: 1%
- Education: 2%
- Fine Arts: 3%
- General Works: 10%
- Geography, Anthropology, Recreation: 0%
- History: 15%
- Language and Literature: 13%
- Law: 0%
- Medicine: 1%
- Military Science: 2%
- Music: 1%
- Naval Science: 2%
- Philosophy, Psychology, Religion: 6%
- Political Science: 2%
- Science: 11%
- Social Sciences: 15%
- Technology: 13%
Patterns in the Overall Online Usage
Lessons Learned

• Improved our understanding of the use of online and print materials made accessible through mass digitization

• Learned from the process about what data is available and where better metrics are needed

• Identified some potential patterns for further study
“The temptation to form premature theories upon insufficient data is the bane of our profession.”

- Sherlock Holmes
Data-based Preservation Decision Making

Jacob Nadal
Preservation Officer
UCLA Library
Preservation Theory and History: Medicine, Zoos and Fortresses

Controversial assertion: *What libraries call preservation is more like conservation at scale, and it’s still not to scale.*

- **20th century preservation was effectively local.** We tried to protect or repair the **items** in the collection
  - Rigid, comprehensive security and environmental standards
  - Fortification of item (library binding, deacidification)
  - Replacement of weak items with hardened versions (library editions, microfilm, facsimiles)

- Libraries are more like zoos than fortresses
- Preservation was trying to deal with public health problems in the metaphorical emergency room
Preservation Theory at UCLA: Public Health, Flood Control & Habitats

- Preservation works from the collection down
- Conservation works from the item up
- At UCLA, one strategy governs both approaches
  - Every activity has a:
    - 1) Method of analysis or evidence-gathering
    - 2) Treatment proposal & outside review
    - 3) Hedge or fail-safe option
- We’re operating a dam or flood control channel, not manning a wall under siege
  - You can make your LA River jokes now... ha, ha...
- As our program matures, the watchword is habitat:
  - Habitats are flexible and adaptable
  - Habitats are sustainable or not, depending on certain pressures
  - Habitats have local versions of global types
The Los Angeles River

The concrete bottom reduces the effectiveness for flood control, creates a bad habitat for wildlife, and ruins its recreational value. All that effort for nothing!


The natural river is less work and functions better. A model worth emulating!
From Theory towards Practice: Habitat, Sure, but Who Lives There?

• UCLA, like all big RLs, has some really shabby books.

• These “brittle books” are frustrating.
  – Repair is not the answer: little structural integrity means they’re irreparable or require “heroic” conservation
  – Reformattting is costly: fragile, poor contrast materials, so scanning has to be careful and high-quality

• And yet, we’re obligated to preserve certain things:
  – Materials with high Los Angelocity
  – Scarce within the UC system, California, or the world
  – Signature collections, future classics, academic emphases

• Everything else, we want you to do for us
  – kthanksnextslide!
So, how does that dam/wall/habitat thing address the problem of lots of individual shabby books, in the context of a globally-intended collection of record?

- Push decisions from the item to the network context
- Holdings review is the first step
  - Holdings data parsed into global (Worldcat), regional (CA/350 miles of zip code 90095), and system (Worldcat Local/NGM)
  - HathiTrust status checked (Portico, (C)LOCKSS, JSTOR to come?)
- These data are placed into a risk assessment model
- Series of automated recommendations are made
**Risk Assessment Model**  
From Candace Yano (UC Berkeley/Ithaka)

<table>
<thead>
<tr>
<th>Initial number of copies</th>
<th>Survival probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>36.6032%</td>
</tr>
<tr>
<td>2</td>
<td>59.8085%</td>
</tr>
<tr>
<td>3</td>
<td>74.5199%</td>
</tr>
<tr>
<td>4</td>
<td>83.8464%</td>
</tr>
<tr>
<td>5</td>
<td>89.7592%</td>
</tr>
<tr>
<td>6</td>
<td>93.5076%</td>
</tr>
<tr>
<td>7</td>
<td>95.8841%</td>
</tr>
<tr>
<td>8</td>
<td>97.3906%</td>
</tr>
<tr>
<td>9</td>
<td>98.3457%</td>
</tr>
<tr>
<td>10</td>
<td>98.9513%</td>
</tr>
<tr>
<td>11</td>
<td>99.3351%</td>
</tr>
<tr>
<td><strong>12</strong></td>
<td><strong>99.5785%</strong></td>
</tr>
<tr>
<td>13</td>
<td>99.7328%</td>
</tr>
<tr>
<td>14</td>
<td>99.8306%</td>
</tr>
<tr>
<td>15</td>
<td>99.8926%</td>
</tr>
<tr>
<td>16</td>
<td>99.9319%</td>
</tr>
<tr>
<td>17</td>
<td>99.9568%</td>
</tr>
<tr>
<td>18</td>
<td>99.9726%</td>
</tr>
<tr>
<td>19</td>
<td>99.9827%</td>
</tr>
<tr>
<td>20</td>
<td>99.9890%</td>
</tr>
<tr>
<td>21</td>
<td>99.9930%</td>
</tr>
<tr>
<td>22</td>
<td>99.9956%</td>
</tr>
<tr>
<td>23</td>
<td>99.9972%</td>
</tr>
<tr>
<td>24</td>
<td>99.9982%</td>
</tr>
<tr>
<td>25</td>
<td>99.9989%</td>
</tr>
<tr>
<td><strong>26</strong></td>
<td><strong>99.9993%</strong></td>
</tr>
</tbody>
</table>

12 copies is where the curve is asymptotic  
26 is derived from past decisions by UCLA
Basic Scenario for Preservation Review

- Three outcomes:
  - Keep if <$12$ global$\ OR <$3$ CA$\ OR [0 UC$\]
  - Withdraw if [$>26$ global$]$
  - Else Review

- Data is collected (point 1) then a proposed treatment gets external review by coll. managers (point 2) and all decisions are hedged by the network (point 3)
  - “Keep” implies that preservation will see to it that the content remains in the collection
  - “Withdraw” really means withdraw
  - “Review” means we need a genuine person to make a decision (and people are both slow and idiosyncratic, so...)
Alternate Scenarios

• Decision making starts with the basic scenario. We’ll fine-tune that as we collect decision data
  • Seeking best match between collection managers decisions and automated indicators
  • After a designated review period, may use a more risk-tolerant scenario to decide on materials lingering in “review” status

• At present, we have a known unknown regarding artifactual value
  • Conservation screens materials and routes to preservation review. Conservators are eagle-eyed about artifactual value
  • Preservation officer reviews all “withdraws” and, for better or worse, yours truly has a mild case of bibliomania
The Hedgerow

#1 OR, Retain First

#2 OR, with Hathi, Retain First

#3 AND, Retain First

#4 AND, with Hathi, Retain First

- **Withdraw**
- **Retain**
- **Review**
The Long Tail

![Graph showing the long tail phenomenon with different categories and data points]
The Los Angeles Triangle

Inside this zone, we have a stewardship obligation, driven by preservation, a general good.

Outside, we have options, driven by an institutional intention.
Acknowledgements and Next Steps

• What made this possible
  • Annie Peterson - summer intern in the UCLA Library preservation office, from UIUC GSLIS. What made it possible
  • Willingness by all to try a “Cynefin” style of work -- gradual sense-making and continuous process improvement

• What would make it easier
  • In-house statistics expertise and research support
  • Longer stretches of uninterrupted time
  • Better serials data - Communal 583 + Local Holdings Records

• What comes next
  • More of the same, to refine and test our process
  • Application to other activities: gifts & exchange, replacements and preservation-driven acquisition, preservation survey & audit
For More Information

ReCAP Data Center (Columbia University)


HathiTrust Digital Library


UCLA Library Preservation Blog

- Jake Nadal and John Riemer “Preservation Actions, MARC 21 Field 583, and Communal Local Holdings in OCLC WorldCat” (CONSER, 2009)
Questions, Comments?

Zack Lane - zl2114@columbia.edu

Helen Look - hlook@umich.edu

Jacob Nadal - jnadal@library.ucla.edu

http://www.oclc.org/research/events/webinars.htm
http://itunes.apple.com/podcast/oclc-research-podcasts-webinars/id284764834