We Don’t Need that Anymore, Exploring the Realities of the Impact of Digitization on Print Usage

Thomas Teper and Vera Vasileva

April 11, Cleveland, OH
We would like to THANK the SPONSOR of the presentation

Morgan & Claypool Publishers
Overview

- Introduction
- Methodology & Problem Statement
- Preliminary Results
- Conclusions
Introduction

“There are two times in a man's life when he should not speculate: when he can't afford it and when he can.”

- Mark Twain, *Following the Equator*, Pudd'nhead Wilson's New Calendar
Methodology & Problem Statement

The project sought to explore:

1. Subject-based differences in the demand for the print titles
2. Difference in items’ demand before and after digitization
3. Difference in demand after digitization for items published before and after 1923
Other questions for eventual evaluation:

• Difference in local (local circ) vs. external demand (ILL/DD)
• Does this difference correspond to how commonly an item is held per holdings registered in OCLC
• Difference in the circulation/ILL data for items sourced from the “large, research university’s” collection and elsewhere
• How did the usage of commercially digitized backfile content influence our assumptions about use of this monographic content?
• Should evidence for user preferences for newly published ebooks inform our decision-making on the management and/or retention of historic print copies?
Primary Challenge

- Voluminous size of the datasets
- Datasets from three different sources
- Datasets did not naturally speak to one another
Pursued Steps to Surmount the Challenge

**Step 1:** Gather datasets

**Step 2:** Import datasets into MS SQL Server database

**Step 3:** Perform data cleaning and manipulation

**Step 4:** Query data for the analysis to answer the research questions
<table>
<thead>
<tr>
<th>Description</th>
<th>Dataset 1</th>
<th>Dataset 2.1</th>
<th>Dataset 2.2</th>
<th>Dataset 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>Record set of every physical item</td>
<td>Original circulations of all physical collection</td>
<td>Currently checked out items</td>
<td>Items digitized from U of I collection</td>
</tr>
<tr>
<td>Content</td>
<td>All items in the print collection</td>
<td>Archive transactions</td>
<td>Current transactions</td>
<td>All digitized items</td>
</tr>
<tr>
<td>Number of records</td>
<td>10.7 million</td>
<td>8,246,410</td>
<td>93,076</td>
<td>751,522</td>
</tr>
<tr>
<td>Source</td>
<td>Voyager catalog</td>
<td>MS Access database</td>
<td>MS Access database</td>
<td>Hathifiles repository</td>
</tr>
</tbody>
</table>
Step 2

Putting the Datasets Together

- Library’s print collection dataset
- Archive transactions dataset
- Current transactions dataset
- Dataset of digitized items (HathiTrust)

MS SQL Server Database
## Step 3  Data Cleaning and Manipulation

### Data Manipulation

- Deduplication of the Library’s print collection dataset
- Assigning subject heading and classification system to a title according to its call number
- Finding common elements to connect the datasets
- Narrowing the datasets

### Data Cleaning

- Format
- Subject heading
- Classification system
- Publication Year
- Digitization Year
- Circulation Year
Narrowing the Datasets
Library’s Print Collection Dataset

- 10,709,114 entire dataset
- 10,601,294 deduplicated
  - 3,185,040 serials
  - 6,909,723 monographs
    - 4,123,726 single-volume monographs
    - 1,181,782 multi-volume monographs
  - 506,531 other

5,305,508

Narrowing the Datasets

Dataset of Digitized Items (HathiTrust)

- **751,522** entire dataset
- **389,794** monographs digitized in 2010-10/2018
- **265,202** monographs digitized in 2010-2016
Narrowing the Datasets
Circulation Dataset
(archive and current transactions)

8,246,410
all archive transactions

93,076
all current transactions

8,339,486
total transactions in 2002-10/2018

8,061,596
total transactions in 2002-2017
Connecting Three Datasets

Library’s Print Collection Dataset

Dataset of Digitized Items (HathiTrust)

Circulation Dataset (archive and current)

connect by

source_bib_num (Bib_ID)
description_h (ENUM)

to obtain Item_ID

connect by
to pull circulations

Bib_ID ENUM

Item_ID

Item_ID
Preliminary Results
RESEARCH QUESTION 1

Are there subject-based differences in the ongoing demand for print resources?

Results indicate that **there are measurable differences in the usage of print resources based on their classifications**
Library’s Print Collection by Classification System

10,601,294 deduplicated dataset

5,603,032 Dewey DC

1,178,936 Library of Congress

321,612 Gov. Docs

1,950,745 not classified yet

1,546,969 local schemas, withdrawn items, errors
Number of Items Used in the Subject-based Analysis

5,603,032 Dewey DC
1,178,936 Library of Congress

6,781,968
Number of Items Circulated in 2002 – 2017

- **5,243,857** total circulations
- **3.3** circ. per item

- **5,176,722** (76%) NOT-Circulated Items
- **1,605,246** (24%) Circulated Items
Average Circulation per Title by Subject Area
RESEARCH QUESTION 2

Is there a measurable difference in demand for these print resources from the periods before and after an item was digitized?

Preliminary results indicate that, there is a decrease in demand for print resources in the periods after their digitization.
Additional Narrowing the Dataset of Digitized Items (HathiTrust)

- 751,522 entire dataset
- 389,794 monographs digitized in 2010-10/2018
- 265,202 monographs digitized in 2010-2016
- 209,398 found Item_ID
- 12.8 % showed evidence of circulation in 2002-2017
- 26,918 circulations
- 42,689 circulations
Average Number of Circulations Per Year Before and After Digitization Year

- Number of Circulations Per Year Before Digitization
- Number of Circulations Per Year After Digitization
RESEARCH QUESTION 3

Is there a difference in demand after digitization for those items that are freely available as full text (most being pre-1923) when compared to those in which copyright or other restrictions limit the digital access?

The results appear to contradict speculation that the availability of full-text access decreases demand on the original item more than those only partially available.

YES
Another Narrowing the Dataset of Digitized Items (HathiTrust)

- 751,522 entire dataset
- 389,794 monographs digitized in 2010-10/2018
- 265,202 monographs digitized in 2010-2016
- 209,398 found Item_ID

206,559 after excluding items with bad Pub.Date

E.g.: 192u, 19uu, 1uuu
Number of Digitized Monographs by Year of Digitization

The graph shows the number of digitized monographs by year of digitization. The lines represent items digitized before 1923 (blue) and items digitized after 1923 (orange). The data is as follows:

- 2010: 5,000 pre-1923, 1,000 post-1923
- 2011: 2,000 pre-1923, 1,000 post-1923
- 2012: 15,000 pre-1923, 30,000 post-1923
- 2013: 10,000 pre-1923, 20,000 post-1923
- 2014: 45,000 pre-1923, 40,000 post-1923
- 2015: 30,000 pre-1923, 25,000 post-1923
- 2016: 20,000 pre-1923, 15,000 post-1923

The data indicates a significant increase in the number of digitized items post-1923, with peaks in 2012 and 2014.
## Overall Statistics

### (Pre- and Post-Digitization)

<table>
<thead>
<tr>
<th></th>
<th>NUMBER OF</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Items</td>
<td>Circulated Items</td>
<td>Circulations</td>
<td>Frequency of Use</td>
</tr>
<tr>
<td>pre-1923</td>
<td>108,836</td>
<td>17,500</td>
<td>27,477</td>
<td>1.57</td>
</tr>
<tr>
<td></td>
<td>52.7 %</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>16.08 %</td>
<td></td>
<td></td>
</tr>
<tr>
<td>post-1923</td>
<td>97,723</td>
<td>8,982</td>
<td>14,513</td>
<td>1.62</td>
</tr>
<tr>
<td></td>
<td>47.3 %</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>9.19 %</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOTAL</td>
<td>206,559</td>
<td>26,482</td>
<td>41,990</td>
<td>~ 1.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Average Number of Circulations Per Year Before and After Digitization Year

pre-1923 Items
Average Number of Circulations Per Year Before and After Digitization Year

post-1923 Items
# Circulations Comparison

<table>
<thead>
<tr>
<th></th>
<th>pre-1923 (108,836 items)</th>
<th>post-1923 (97,723 items)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>before digitization</td>
<td>after digitization</td>
</tr>
<tr>
<td><strong>Total circulations</strong></td>
<td>25,923</td>
<td>1,522</td>
</tr>
<tr>
<td></td>
<td>13,836</td>
<td>677</td>
</tr>
<tr>
<td><strong>Circulation per Item</strong></td>
<td>0.238</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>0.142</td>
<td>0.0069</td>
</tr>
</tbody>
</table>

Before digitization: \( \frac{0.238}{0.142} \approx 1.68 \)

After digitization: \( \frac{0.014}{0.0069} \approx 2.02 \)
Circulation of Digitized pre-1923 Items by Subject Area

- 800 - Literature
- 900 - History & geography
- 300 - Social sciences
- 600 - Technology
- 500 - Science
- 200 - Religion
- 700 - Arts & recreation
- 000 - CS, information & general works
- 400 - Language
- 100 - Philosophy & psychology
- K - Law
- P - Language and literature
- J - Political science

Subject Areas and Circulation:
- P - Language and literature, 36.7%
- 100 - Philosophy & psychology, 25.0%
- 800 - Literature, 15.5%
- 300 - Social sciences, 14.0%
- 500 - Science, 15.4%
- 400 - Language, 13.3%
- 600 - Technology, 8.4%
- J - Political science, 6.7%
- 000 - CS, information & general works, 13.0%
- 700 - Arts & recreation, 22.4%
- 200 - Religion, 19.8%
- 900 - History & geography, 22.6%
Circulation of Digitized post-1923 Items by Subject Area

900 - History & geography, 48.2%
800 - Literature, 24.8%
200 - Religion, 35.7%
P - Language and literature, 23.3%

100 - Philosophy & psychology, 30.2%
400 - Language, 28.5%
700 - Arts & recreation, 26.5%

500 - Science, 15.1%
000 - CS, information & general works, 13.0%
D - World history and history of Europe, ..., 11.3%
300 - Social sciences, 11.0%
K - Law, 8.1%
600 - Technology, 8.0%

- 300 - Social sciences
- 600 - Technology
- 900 - History & geography
- 500 - Science
- 800 - Literature
- 700 - Arts & recreation
- 000 - CS, information & general works
- 200 - Religion
- 400 - Language
- D - World history and history of Europe, ...
- 100 - Philosophy & psychology
- K - Law
- P - Language and literature
Conclusions...

• There are differences in demand by subject area.
• There is a decline in average annual use after digitization.
• This decline occurs regardless of copyright status (at least when defined by year of publication).
  – However:
    • Post digitization, pre-1923 items circulated more frequently than post-1923 items.
Questions?

- **Thomas H. Teper**, Associate University Librarian for Collections and Technical Services and Associate Dean of Libraries; Chair, Military Education Council (2014 – 2020); Associate Professor
  University of Illinois at Urbana-Champaign
  246G Main Library
  1408 West Gregory Drive
  Urbana, IL  61801
  email:  tteper@illinois.edu

- **Vera Vasileva**, MSLIS’19 Candidate, School of Information Sciences; Graduate Assistant, Office of Collections and Technical Services at the University Library, University of Illinois at Urbana-Champaign
  email:  verav2@Illinois.edu