Robert H. McDonald | @mcdonald
Associate Dean for Library Technologies
Deputy Director Data to Insight Center (D2I)
Indiana University

Elephant in the Room: Scaling Storage for the HathiTrust Research Center

@hathitrust

PASIG 2015
#PASIG2015
UC San Diego
March 12, 2015
Mission of the HT Research Center

- Research arm of HathiTrust
- Established: July, 2011
- Collaborative center: Indiana University & University of Illinois
- Mission: Enable researchers world-wide to accomplish tera-scale text data-mining and analysis
  - Develop cyberinfrastructure to enable HPC access to the HathiTrust Digital Library
  - Develop cutting-edge software tools for processing, analyzing text
  - Develop translational tools and data that can be used to enhance HathiTrust Digital Library services to users
HathiTrust and HTRC

Board of Governors
- Executive Committee
- Executive Director

HathiTrust

HathiTrust Research Center

University of Illinois

Indiana University

University of Michigan
Working with HTRC Staff

**Scholarly Commons** ➔ Workshops, tutorials, and guidance for using HTRC

**Advanced Collaborative Support** ➔ One-on-one research support provided through a competitive awards process

**Advanced Research** ➔ Collaborative research partnership with HTRC
HathiTrust “Wow” Numbers

- 13,284,163 total volumes
- 6,742,394 book titles
- 352,534 serial titles
- 4,649,457,050 pages
- 595 terabytes
- 157 miles
- 10,793 tons
- 4,979,599 volumes in the public domain
Non-Consumptive Research Paradigm

• No action or set of actions on part of users, either acting alone or in cooperation with other users over duration of one or multiple sessions can result in sufficient information gathered from collection of copyrighted works to reassemble pages from collection.

• Definition disallows collusion between users, or accumulation of material over time. Differentiates human researcher from proxy which is not a user. Users are human beings.
<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 0</td>
<td>Derived Factual Data or Supplementary Bibliographic Data; Publicly available in US and worldwide (bulk download permitted)</td>
</tr>
<tr>
<td>Level 1A</td>
<td>Page text or Page images that are in public domain in US only and not subject to third party restrictions; Publicly available to everyone in the US (no bulk download)</td>
</tr>
<tr>
<td>Level 1B</td>
<td>Page text or Page images that are in the public domain worldwide and not subject to third party restrictions; Publicly available to everyone worldwide (no bulk download)</td>
</tr>
<tr>
<td>Level 1C</td>
<td>Primary Bibliographic Metadata; Publicly available in US and worldwide (no bulk download)</td>
</tr>
<tr>
<td>Level 2A</td>
<td>Page text or Page images that are in public domain in US only and are subject to third party restrictions; Publicly available to everyone in the US (no download)</td>
</tr>
<tr>
<td>Level 2B</td>
<td>Page text or Page images that are in public domain worldwide and are subject to third party restrictions; Publicly available to everyone worldwide (no download)</td>
</tr>
<tr>
<td>Level 3</td>
<td>Restricted in-copyright data; may or may not be subject to additional third-party restrictions (no download)</td>
</tr>
</tbody>
</table>
Working with HTRC Tools

Get started at: https://htrc2.pti.indiana.edu/

Build Worksets

Execute Algorithms

Visualize Term Frequency

http://sandbox.htrc.illinois.edu/bookworm/
HTRC Shared IU Systems

Data Systems
• Data Capacitor II (Lustre)
• NetApp NFS (18 TB)
  – GPFS/Data Direct Networks (28 TB)

Current Storage (18 TB > 30 TB)
• MARC – 15 GB
• R – 5 TB
• iPython – 5 TB
• Bookworm – 5 TB
• Public Domain BW – 3 TB
• OCR Data – 5 TB
• OCR Index – 2.3 TB
• Audit Logs – 44 GB
• User created MD – 15 GB
• Blacklight – 20 GB
• IU Pairtree – 2.65. TB

Compute Systems
• KARST (high-throughput cluster)
• Big Red 2 (Cray XE6/XK7)
Want More HTRC?

3rd Annual HTRC UnCamp!
March 30-31, 2015 in Ann Arbor, Michigan

DH 2015
June 29-3 July, 2015 in Sydney, Australia

http://www.hathitrust.org/htrc_uncamp2015
Thank You

- This presentation was made possible with content provided by many HTRC colleagues John Unsworth, J. Stephen Downie, Beth Plale, Beth Namachchivaya, Dirk Herr-Hoyman, Milnda Pathirage, Samitha Liyanage, Miao Chen, Guangchen Ruan, Jiaan Zeng, Loretta Auvil, Boris Capitanu, and many others...
- The HTRC Non-Consumptive Research Grant was graciously funded by the Alfred P. Sloan Foundation.
- THE HTRC WCSA grant is graciously funded by the Andrew W. Mellon Foundation.
- HTRC - http://www.hathitrust.org/htrc
- IU D2I Center - http://d2i.indiana.edu/
- UIUC GSLIS - http://www.lis.illinois.edu/