HATHI TRUST
RESEARCH CENTER

HathiTrust Research Center: Technical Challenges

OSP Workshop| 4.June.14

Presented by Miao Chen

Data To Insight Center, Indiana University

INDIANA UNIVERSITY

RESEARCH CENTER

ILLINOIS

Tweet us –
# dlbb
@HathiTrust #HTRC

HathiTrust Digital Library

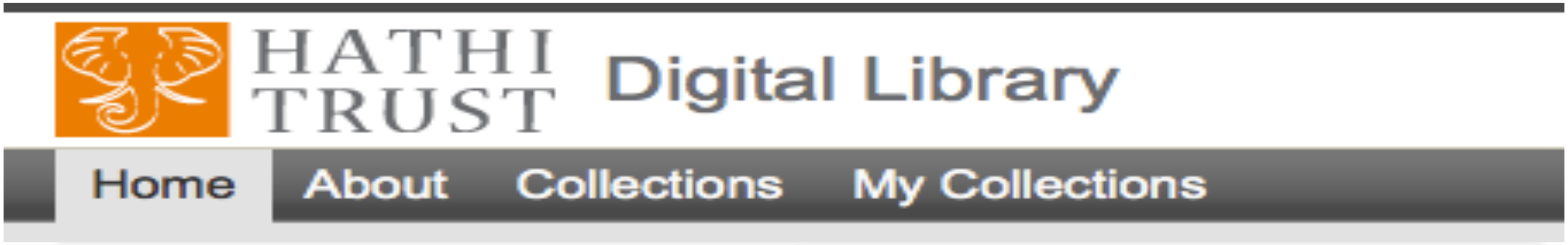
- HathiTrust is a partnership of academic & research institutions, offering a collection of millions of titles digitized from libraries around the world.
  - – Founding members of HathiTrust along with University of Michigan are Indiana University, University of California, and University of Virginia

http://www.hathitrust.org

→ Distinguished from

RESEARCH CENTER

http://www.hathitrust.org/htrc



Currently Digitized (by 6/4/2014)

- 11,167,882 total volumes
- 5,814,665 book titles
- 291,945 serial titles
- 3,908,758,700 pages
- 501 terabytes
- 132 miles
- 9,074 tons
- 3,773,881 volumes(~34% of total) in the public domain

http://www.hathitrust.org/statistics_info

→ HathiTrust repository is a latent goldmine for text mining analysis, analysis of large-scale corpi through computational tools, and time-based analysis

→Restricted nature of HT content suggests need for new forms of access that preserve intimate nature of research investigation while honoring restrictions

→ Paradigm: computation takes place close to the data

Mission of HT Research Center

- Research arm of HathiTrust
- Goal: enable researchers world-wide to carry out computational investigation of HT repository through
  - Develop model for access: the ‘workset’
  - Develop tools that facilitate research by digital humanities and informatics communities
  - Develop secure cyberinfrastructure that allows computational investigation of entire copyrighted and public domain HathiTrust repository
- Established: July, 2011
- Collaborative effort of Indiana University, University of Illinois, and HathiTrust

Technical Challenges

- Computational access for non-consumptive research
- Version control of volumes
- Linkage with external resources
- Data representation
- Machine learning on large-scale text data
- Blurry boundary between metadata and feature of text
- Provenance of scholar's experiments
- Satisfying need of small-scale and large-scale runs
- OCR errors and cleanup