HathiTrust Research Center
The Fast Version

Robert H. McDonald | @mcdonald
Executive Committee-HathiTrust Research Center (HTRC)
Deputy Director-Data to Insight Center
Associate Dean-University Libraries
Indiana University

INDIANA UNIVERSITY

RESEARCH CENTER

ILLINOIS

HTRC Mission

The HathiTrust Research Center (HTRC) is a collaborative research center launched jointly by Indiana University and the University of Illinois to act as the public-facing research arm of the massive HathiTrust Digital Library.

The HTRC is mandated to help researchers from around the world surmount the difficulties associated with processing and analyzing terascale amounts of digital text. Thus, the scholarly developers at HTRC work to develop cutting-edge software tools and cyberinfrastructure to enable advanced computational access to the growing digital record of human knowledge. HTRC began its efforts July 2011.

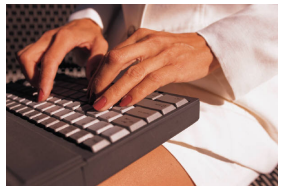
HTRC Non-Consumptive Research Paradigm

- *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.

HTRC Current Infrastructure

- Servers
  - 14 production-level quad-core servers (virtual machines)
    - 16 – 32GB of memory
    - 250 – 500GB of local disk each
  - 6-node Cassandra cluster for volume store
  - Ingest service and secure Data API access point
- Storage (IU University Infrastructure)
  - 13TB of 15,000 RPM SAS disk storage
  - Increase up to 17TB by end of 2012
  - 500TB available in late year 2-year 3

HTRC Architecture



Portal Access

Blacklight

Agent

Job Submission

Collection building

Direct programmatic access (by programs running on HTRC machines)

Security (OAuth2)

Data API access interface

Solr Proxy

Audit

Registry (WSO2)

Algorithms

Meandre Workflows

Result Sets

Collections

Cassandra cluster volume store

Solr index

Compute resources

Storage resources

HTRC Next Steps

- Phase 2 Begins Jan 2013….
- Thanks to

ALFRED P. SLOAN FOUNDATION

Contact Information

- Robert H. McDonald
  - Email – robert@indiana.edu
  - Chat – rhmcdonald on googletalk | skype
  - Twitter - @mcdonald
  - Blog – http://www.rmcdonald.net
  - Twitter Hashtag: #HTRC12
  - Web: http:www.hathitrust.org/htrc

http://slidesha.re/QCOrIX