HTRC: Transition to Outreach Phase II

May 01, 2012 Urbana, Illinois
<table>
<thead>
<tr>
<th>Time</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00 – 9:45</td>
<td>Introduction and Setting Stage for Phase II</td>
</tr>
<tr>
<td>9:45 – 10:45</td>
<td>Roadmap to non-consumptive research</td>
</tr>
<tr>
<td>10:45 – 11:00</td>
<td>break</td>
</tr>
<tr>
<td>11:00 – 12:00</td>
<td>HTRC and HathiTrust</td>
</tr>
<tr>
<td>12:00 – 12:30</td>
<td>HTRC ExecMgt Closed meeting</td>
</tr>
<tr>
<td>12:30 – 2:00</td>
<td>Workset: definition; conceptual and technical implications</td>
</tr>
<tr>
<td>2:00 – 2:15</td>
<td>Break</td>
</tr>
<tr>
<td>2:15 – 4:00</td>
<td>Outreach</td>
</tr>
</tbody>
</table>
Attendees

• HTRC ExMgt
  • Beth Plale, IU
  • J. Stephen Downie, UIUC
  • Beth Sandore, UIUC
  • John Unsworth, Brandies
  • Robert McDonald, IU
• Ted Underwood, English, UIUC
• Inna Kouper, CLIR Fellow, IU

• Yiming Sun, HTRC lead architect
• Tim Cole, UIUC Library
• Miao Chen, HTRC Asst Dir for Outreach and Education
• Megan Finn Senseney, UIUC
• Jeremy York, HathiTrust (remote)
Recent News

• HathiTrust Research Center release of software services stack Mar 31, 2013:

• New tools available to mine world's largest digital repository of books. This week the HathiTrust Research Center (HTRC) announced the availability of data mining and analytics tools for the HathiTrust Digital Library, a collection of digital texts from over 70 research libraries around the world. The new tools provide a much-needed entry point to large-scale analysis of HathiTrust's contents.

• 'Getting started FAQ' is here: http://bit.ly/XkZKev
What’s Next

• **HTRC Face-to-Face:** May 01, 2013, Urbana, IL. Includes HTRC ExMgt team, technical leads, and for a portion, Jeremy York and John Wilkin. Purpose is next year planning.

• **Non-consumptive Research:** have technical and strategic roadmap to get us to Fall 2013 goal of security copy of copyrighted portion of corpus. Call with HTRC AB members was first step. HTRC software and services release Mar 31, 2013 has version of Sloan funded HTRC Sloan Cloud in its development version. HTRC Sloan Cloud will see two more releases with increased security refinements, Summer 2013 and Fall 2013.

• **Technical and conceptual:** HTRC will be developing out the notion of the workset as both a conceptual and technical notion for using HTRC.
Pending and Current Funding

• Pending:
  • Workset Creation for Scholarly Analysis, Mellon Foundation, pending Mar 2013

• Current:
  • Sloan Foundation, Data Capsule for Non-Consumptive Research, 7/2011 – 6/2014
  • Data To Insight internal funding
  • UIUC internal funding
Engagement history with HTRC and advisory boards

• Dec 14 and 20, 2011: Virtual meeting with HTRC advisory board
• May 2011: Virtual meeting with HathiTrust SAB
• Sep 2012: HTRC UnCamp
• Nov 2012: Meeting with select HTRC Advisory Board to discuss strategy for non-consumptive research
• April 2013: notified HTRC AB of software release
Community Engagement

• HTRC-usergroup-l list: users of HTRC system and services. 26 people on list. Engagement from Texas A&M Univ, McGill, Harvard, U Wisconsin, U Chicago, Notre Dame, New York University, …

• 2nd HTRC UnCamp 2013 planned for September 2013, to be held in Urbana, Illinois.

• Planned local outreach workshop: Fall 2013 at IU in collaboration with the Catapult Center for Digital Humanities and Computational Analysis (http://www.indiana.edu/~catapult/)
Community Engagement, cont.

• Scholarly commons at UIUC. Univ library and GLIS establish commons as outreach point. 4 scholars across campus. Text data mining part of their research. Steven invited to talk about HTRC. Focused on identify opportunities for innovation. Already publicizing access to HTRC data.

• "Humanities without walls" CIC effort; Mellon funded. Be purposeful in talking with CIC about their engagement on digital humanities.
HTRC Phase II: Objectives

• Outreach: plan and budget for ‘13-’14 AY
• Software development: Streamline development effort. Priority on:
  • *User-driven requirements: track, prioritize*
  • *Bugs*
  • *Simplification/ease of management*
  • *HTRC Sloan Cloud for non-consumptive research*
• Improved funding efforts – stronger position
• Improved reporting / tracking
### HTRC Tech Stack Deployment Timeline

**Deliver: Mar 31, 2013**
- **Sandbox stack (resides at UIUC):** non-google corpus (250,000 volumes), open access.
- **Production stack (resides at IU):** v0.5 in place. Uses Oauth security. Public domain corpus. Shares Cassandra/Solr with dev stack. Minimal compute resources available.
- **Development stack (resides at IU):** shares Cassandra/Solr with prod stack. Supports v0.1 of HTRC Sloan Cloud for non-consumptive support.

**Deliver: June 30, 2013**
- **Sandbox stack (at UIUC):** v1.0 stack but against non-google corpus
- **Production stack (at IU):** v1.0 reflects extensive testing. Oauth for security. Public domain corpus. Share Cassandra/Solr with dev stack. Support for parallel execution.
- **Development stack (at IU):** share Cassandra/Solr with prod stack. New services. V0.2 of Sloan non-consumptive support. Begin dev for InCommon and auditing.
- **Documentation:** Transparent documentation on security

**Deliver: Sep 30, 2013**
- **Sandbox stack (at UIUC):** v1.5; against non-google corpus
- **Production stack (at IU):** v1.5. Supports InCommon in anticipation of copyright works. Public domain corpus. Separate Cassandra/Solr; public domain corpus
- **Development stack (at IU):** InCommon, auditing, and v1.0 of Sloan non-consumptive support. Security audit on development stack; verify ready for copyright materials.

**Deliver: Nov 30, 2013**
- **Sandbox stack:** retire (?)
- **Production stack (at IU or UIUC):** v2.0. Supports InCommon in anticipation of copyright works. Public domain corpus. Separate Cassandra and Solr for public domain corpus.
- **Development stack (at IU or UIUC):** dev stack ready for copyright materials.
Non-consumptive research
Categories of algorithms. Can fair use be determined based on categorization of algorithm? Or is all computational use fair use?
Categories of algorithms. Can fair use be determined based on categorization of algorithm? Or is all computational use fair use?

Believe now all are fair use; search exception if no review of results done. Nuance in interaction mode.
Non-consumptive research

- What is nature of text analytics? Volume list as input, compute, results set as output or more interactive? Intermediate results could be a page, metadata. Issue is in handling corpus at scale. Reading portions of text, marking up portions of text. Libraries exercise fair use analysis. Could be challenged that copying corpus from one place to another is not fair use. Human in loop: Review new jobs, algorithms.

- Security check: Hack-a-thon: assemble group who’s job it is to break through security barriers of HTRC Sloan Cloud

- HTRC as pure research center: keeping corpus and metadata in original form is for HT. HTRC has freedom to let community experiment with enhanced OCR and augmented metadata.

- Services that add value (facilitate GUI such as http://imlsdcc.grainger.uiuc.edu/):
  - Mining metadata of a collection; used to describe a collection more fully. Provides context information about collections.
  - Error correction in the OCR. Would benefit every user. Adding classifiers to metadata. We do things to collection that aren’t in what we got from HT. Also factors into issues of copyright. This value add to the corpus is an original piece of work that works well in defending our use.
  - Book search is a hard problem. Info about what is on each page, and relationships between pages.
Steps to non-consumptive research

• Philosophy: we don't push anything back to HathiTrust. HT gets periodic updates. We have earlier copy that we will change. We have one version of corpus. We can take updates, run OCR correction, etc on new version.

• Two applications that define:
  • Six degrees of Francis Bacon; metadata enhancement and OCR correction (remediated text). Music score detection across corpus images; find where music exists. Extend to looking for other things: faces, tables, maps (DeRoure)
  • Ted Underwood: doing genre mapping for 18th - 20th century books. Generates metadata that would be of interest. Willing to share.

• Notes from call with HTRC AB
  • https://docs.google.com/document/d/1WwQx3oRX1ys-rz2xh_g9JlA2Sfx0kTw8RX-oHidhg/edit

• Technical
  • HTRC Sloan Cloud leaving ‘breadcrumbs trail’ of security for transparency in reporting and easier security assessment. Wiki under development <URL>
  • Security “certification” and hack-a-thon
HTRC – HT discussion
HT – HTRC talking points

• Help us with primary objectives: outreach, fund raising.
• Uncertainty in HT governance; impact on HTRC
• Assure stronger reporting
HT-HTRC: Non-consumptive

• Prepare document on revenue sharing policies: members of HathiTrust get more benefit: aka, more computation than non-HT members. Part of funding goes to research center. (Unsworth)

• Prepare materials to make non-consumptive argument: Three part: 1) we have 2 proposals in hand, both of which need access to copyrighted content. 2) HTRC-Sloan-Cloud: document on measures in place. Architecture (aka, vms, networking, storage), Security assessment done by external party. Description of use. This is how we mean to do non-consumptive research. (Plale, Unsworth)

• Strategy: Jack Bernard review is first step. HTEC gets review. Then pass past counsels at IU and UIUC. Google will likely need to be involved.
HT-HTRC: Community/Outreach

• Move community presence to HT. Write up document: what services are located where (bug tracking, portal, blacklight, community presence …) (M. Chen)

• To build community/use cases: Go back to those who gave input to HTRC CFP (M. Chen)

• HT willing to help us by engaging people; would put out survey

• What are benefits/features of HTRC (educational materials) (M. Chen)
Workset: conceptual and technical implications
Workset: conceptual and technical implications

• Workset Defn: set of pointers to all or part of any number of items in the HT corpus and external to the corpus

• HT has “collection builder” built manually then saved. People in text analytics need to gather objects (10,000) in large numbers – can’t necessarily be built manually (augment workset by learning from hand-built set). Currently have crude notion of collection as list of volume IDs. Reimagine what objects are: could be pictures on a page. Deconstructing the page, the volume. Notions of page, chapter. Ability to point at, and move around. Aggregations of things within works. Able to point to ‘things’ that are also outside HTRC: e.g. sentiment label stored in semantic web. This workset (similar to research object) is then passed in for computation.
Workset: conceptual and technical implications

• Subsets constructed from the parent workset by different researchers. 19th century, poems, just poems themselves (verse portion) and set of document IDs.

• Worksets could be moveable; could be worked on outside the non-consumptive environment. Workset in this case may contain metadata but data. Workset itself needs describing. Potential public work sets: genre classification, language, removing running headers.

• HTRC: bring in LOC and images. What is the addressable entity? Addressability below page level not needed. Addressability below volume level may motivate move from Cassandra to mongoDB.

• Provenance of analysis process for reproducibility (e.g. Karma)
Workset: conceptual and technical implications


• What are persistent services we could make avail to users? E.g, Popular Passages, Martin Wattenburg. Passage from Shakespeare gets quoted in these other literature works. (tu)

• Workset: metadata pertaining to worksets exportable to Zotero as collection (Monk did this). Zotero has broad user base, could tie in. HTRC philosophy: be part of broader DH community software/services ecosystem (ju)

• Open Annotation –rdf style annotation representation tool.
Outreach Discussion
Outreach: UnCamp objectives and plans

• Email initial plan to HTRC AB for feedback.
• Provost gives opening remarks
• Day 1: higher level; Day 2: smaller group, more focused hands-on
• Situate project (town hall; but as part of kickoff): what happened last time/what we’re doing. For things that are hard (haven’t been done), put back to community and have them prioritize
• New technical functionality that takes advantage of pre-computed metadata:
  • Access to precalculated feature sets (word counts, lines in page, # lines that being with capital letter) at a volume level; select workset (classification algorithm) do API for feature counts (could be precalculated). Effectively no limit on size of workset
  • Language, gender of author
Outreach: UnCamp objectives and plans

• Gather system requirements
• User-side research of HT/HTRC: CIRSS Carole Palmer expressed interest in
• Workset: Gather interest in workset
• Scholarly commons office hours: intake form contains use case. Use to capture needs; determine gap between what we were able to provide and what was needed. Sara Shreeves willing to help.
• HTRC in the classroom: number of attendees of last uncamp were interested in this topic
• Keynote: Six degrees of Francis Bacon (CMU); Matt Wilkin (ND; gives good presentation; interested in non-consumptive);
Outreach: UnCamp objectives and plans

• Panel on Funding: Josh Greenburg, Sloan Foundation; Don Waters, Mellon; NSF SBE, Brett Bobley NEH ODH, Mara Marks, IMLS. Myron Gutmann (NSF) or replacement. [be prepared with material to describe process of partnering with us]

• Panel of Experts: HT BOG (B. Schotlander, Wendy Louge, Laine Farley, CDL) and HTRC advisory group. Back of Book Indexes, Edie Rasmussen; Laura Mandel, Texas A&M. Ben Bennel Google (?)

• Invite Proquest contact, Elsevier

• Build in break times that are long enough for discussions (30 min)

• Balance priorities: Uncamp new features; non-consumptive progress; loc +images added