Two Projects, One Challenge:
Common research data issues in MIREDX and HTRC
Presented by
J. Stephen Downie
University of Illinois at Urbana-Champaign
Acknowledgements

• Most of today’s slides are directly drawn (aka copied) from the slides presented at the HTRC UnCamp events in Bloomington, Indiana and Champaign, Illinois.

• Today’s talk summarizes four days of excellent presentations and demonstrations!

• We thank the HTRC team and the UnCamp presenters for the use of their very informative slides.
Agenda

• Introduce MIREX
• Introducing the HathiTrust
• Non-Consumptive Research
  – MIREX and NEMA
  – HTRC Software Architecture
• Workset Creation for Scholarly Analysis
• Next Steps
In the Beginning...

• Work began on MIREX in Bloomington, Indiana in 2001
  – The “Indiana Manifesto”
• 2001-2003
  – Fact-finding meetings, planning meetings, and workshops funded by Mellon and NSF
  – Large-scale funding from NSF and Mellon 2003
• Audio Description Contest at ISMIR 2004
• MIREX first run at ISMIR 2005 in London
MIREX Model

• Based upon the TREC approach:
  – Standardized queries/tasks
  – Standardized collections
  – Standardized evaluations of results

• Not like TREC with regard to distributing data collections to participants
  – Music copyright issues, ground-truth issues, overfitting issues
IMIRSEL: First Principles

1. Security for the music materials
2. Accessibility for international, domestic and internal researchers
3. Sufficient computing and storage infrastructure for the computationally- and data-intensive MIR/MDL techniques examined
MIREX Overview

• Began as MIREX in 2005
• Tasks defined by community debate
• Data sets collected and/or donated
• Participants submit code to IMIRSEL
• Code rarely works first try 😊
• Huge labour consumption getting programmes to work
• Meet at ISMIR to discuss results
MIREX Model

• Based upon the TREC approach:
  – Standardized queries/tasks
  – Standardized collections
  – Standardized evaluations of results

• Not like TREC with regard to distributing data collections to participants
  – Music copyright issues, ground-truth issues, overfitting issues
IMIRSEL: First Principles

1. Security for the music materials
2. Accessibility for international, domestic and internal researchers
3. Sufficient computing and storage infrastructure for the computationally- and data-intensive MIR/MDL techniques examined
IMIRSEL Model

Legend:
- Super-Bandwidth I/O Channel
- NCSA Music Data Secure Zone
- Command-Control/Derived Data traffic via Internet
- Connection to International MIR Grid
MIREX Participation

MIREX 2012: 109 participants from 20 countries
### MIREX Participation

<table>
<thead>
<tr>
<th></th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Datasets</strong></td>
<td>10</td>
<td>13</td>
<td>12</td>
<td>18</td>
<td>26</td>
<td>31</td>
<td>32</td>
<td>35</td>
<td>37</td>
</tr>
<tr>
<td><strong>Individuals</strong></td>
<td>82</td>
<td>50</td>
<td>73</td>
<td>84</td>
<td>138</td>
<td>152</td>
<td>156</td>
<td>109</td>
<td>116</td>
</tr>
<tr>
<td><strong>Runs</strong></td>
<td>86</td>
<td>92</td>
<td>122</td>
<td>169</td>
<td>289</td>
<td>337</td>
<td>312</td>
<td>302</td>
<td>328</td>
</tr>
</tbody>
</table>
MIREX 2013

• 116 researchers
• More than 29 countries
• 37 datasets
• 24 tasks
• 328 completed runs
<table>
<thead>
<tr>
<th>MIREX 2013 TASKS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio Artist Identification</td>
</tr>
<tr>
<td>Audio Beat Tracking</td>
</tr>
<tr>
<td>Audio Chord Detection</td>
</tr>
<tr>
<td>Audio Classical Composer ID</td>
</tr>
<tr>
<td>Audio Cover Song Identification</td>
</tr>
<tr>
<td>Audio Drum Detection</td>
</tr>
<tr>
<td>Audio Genre Classification</td>
</tr>
<tr>
<td>Audio Key Finding</td>
</tr>
<tr>
<td>Audio Melody Extraction</td>
</tr>
<tr>
<td>Audio Mood Classification</td>
</tr>
<tr>
<td>Audio Music Similarity</td>
</tr>
<tr>
<td>Discovery of Repeated Themes &amp; Sections</td>
</tr>
</tbody>
</table>
# MIREX Participation

<table>
<thead>
<tr>
<th></th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
</tr>
</thead>
<tbody>
<tr>
<td>Datasets</td>
<td>10</td>
<td>13</td>
<td>12</td>
<td>18</td>
<td>26</td>
<td>31</td>
<td>32</td>
<td>35</td>
<td>37</td>
</tr>
<tr>
<td>Individuals</td>
<td>82</td>
<td>50</td>
<td>73</td>
<td>84</td>
<td>138</td>
<td>152</td>
<td>156</td>
<td>109</td>
<td>116</td>
</tr>
<tr>
<td>Runs</td>
<td>86</td>
<td>92</td>
<td>122</td>
<td>169</td>
<td>289</td>
<td>337</td>
<td>312</td>
<td>302</td>
<td>328</td>
</tr>
</tbody>
</table>

**Total Runs: 2037!**
Introducing the HathiTrust
<table>
<thead>
<tr>
<th>Universities in North America</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allegheny College</td>
</tr>
<tr>
<td>Arizona State University</td>
</tr>
<tr>
<td>Baylor University</td>
</tr>
<tr>
<td>Boston College</td>
</tr>
<tr>
<td>Boston University</td>
</tr>
<tr>
<td>Brown University</td>
</tr>
<tr>
<td>California Digital Library</td>
</tr>
<tr>
<td>Colby College</td>
</tr>
<tr>
<td>Columbia University</td>
</tr>
<tr>
<td>Cornell University</td>
</tr>
<tr>
<td>Dartmouth College</td>
</tr>
<tr>
<td>Duke University</td>
</tr>
<tr>
<td>Emory University</td>
</tr>
<tr>
<td>Florida State University</td>
</tr>
<tr>
<td>Getty Research Institute</td>
</tr>
<tr>
<td>Harvard University Library</td>
</tr>
<tr>
<td>Indiana University</td>
</tr>
<tr>
<td>Johns Hopkins University</td>
</tr>
<tr>
<td>Lafayette College</td>
</tr>
<tr>
<td>Library of Congress</td>
</tr>
<tr>
<td>Massachusetts Institute of</td>
</tr>
<tr>
<td>Technology</td>
</tr>
<tr>
<td>McGill University</td>
</tr>
<tr>
<td>Michigan State University</td>
</tr>
<tr>
<td>New York Public Library</td>
</tr>
<tr>
<td>New York University</td>
</tr>
<tr>
<td>North Carolina Central</td>
</tr>
<tr>
<td>University</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Universities in Europe</th>
</tr>
</thead>
<tbody>
<tr>
<td>University of Alberta</td>
</tr>
<tr>
<td>University of British Columbia</td>
</tr>
<tr>
<td>University of Arizona</td>
</tr>
<tr>
<td>University of Calgary</td>
</tr>
<tr>
<td>University of California</td>
</tr>
<tr>
<td>Berkeley</td>
</tr>
<tr>
<td>Davis</td>
</tr>
<tr>
<td>Irvine</td>
</tr>
<tr>
<td>Los Angeles</td>
</tr>
<tr>
<td>Merced</td>
</tr>
<tr>
<td>Riverside</td>
</tr>
<tr>
<td>San Diego</td>
</tr>
<tr>
<td>San Francisco</td>
</tr>
<tr>
<td>Santa Barbara</td>
</tr>
<tr>
<td>Santa Cruz</td>
</tr>
<tr>
<td>The University of Chicago</td>
</tr>
<tr>
<td>University of Connecticut</td>
</tr>
<tr>
<td>University of Delaware</td>
</tr>
<tr>
<td>University of Florida</td>
</tr>
<tr>
<td>University of Houston</td>
</tr>
<tr>
<td>University of Illinois</td>
</tr>
<tr>
<td>University of Illinois at Chicago</td>
</tr>
<tr>
<td>The University of Iowa</td>
</tr>
<tr>
<td>University of Maryland</td>
</tr>
<tr>
<td>University of Massachusetts</td>
</tr>
<tr>
<td>University of Miami</td>
</tr>
<tr>
<td>University of Michigan</td>
</tr>
<tr>
<td>University of Minnesota</td>
</tr>
<tr>
<td>University of Missouri</td>
</tr>
<tr>
<td>University of Nebraska-Lincoln</td>
</tr>
<tr>
<td>The University of North</td>
</tr>
<tr>
<td>Carolina at Chapel Hill</td>
</tr>
<tr>
<td>University of Notre Dame</td>
</tr>
<tr>
<td>University of Oklahoma</td>
</tr>
<tr>
<td>University of Pennsylvania</td>
</tr>
<tr>
<td>University of Pittsburgh</td>
</tr>
<tr>
<td>University of Queensland</td>
</tr>
<tr>
<td>University of Tennessee,Knoxville</td>
</tr>
<tr>
<td>University of Utah</td>
</tr>
<tr>
<td>University of Virginia</td>
</tr>
<tr>
<td>University of Washington</td>
</tr>
<tr>
<td>University of Wisconsin-Madison</td>
</tr>
<tr>
<td>Utah State University</td>
</tr>
<tr>
<td>Wake Forest University</td>
</tr>
<tr>
<td>Washington University</td>
</tr>
<tr>
<td>Yale University Library</td>
</tr>
</tbody>
</table>
Mission

To contribute to the common good by collecting, organizing, preserving, communicating, and sharing the record of human knowledge
HathiTrust “Wow” Numbers

- 10,924,244 total volumes
- 5,719,252 book titles
- 285,776 serial titles
- 3,823,485,400 pages
- 490 terabytes
- 129 miles
- 8,876 tons
- 3,565,657 volumes (~33% of total) in the public domain
Content Distribution

- In-copyright or undetermined: 70%
- "Public Domain": 30%
- Public Domain (worldwide): 15%
- U.S. Federal Government Documents (worldwide): 4%
- Public Domain (US): 10%
- Open Access: .1%
- Creative Commons: .01%

"Public Domain" 30%

Public Domain (worldwide) 15%

U.S. Federal Government Documents (worldwide) 4%

Public Domain (US) 10%

Open Access .1%

Creative Commons .01%
Google PD Research Collection

• Public Domain Materials of the HatihTrust
  – 2,592,097 Volumes
  – Gigabytes
    • 2.3 TB in raw OCR’d text
    • 3.7 TB of managed OCR’d text
    • 1.85 TB solr Index
  – Monthly Updates
    • And irregular data ‘take down’ requests
Non-Consumptive Research Model
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.
Non-Consumptive Research Paradigm

Bring the COMPUTATION to the DATA!
An External Classification Algorithm
Networked Environment for Music Analysis

NEMA DIY Interface
Networked Environment for Music Analysis

NEMA DIY Interface
Evaluation Reports

MIREX 2010: Audio Chord Description - MIREX09 Dataset

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task ID</td>
<td>17</td>
</tr>
<tr>
<td>Task Name</td>
<td>MIREX2010: Audio Chord Description - MIREX09 Dataset</td>
</tr>
</tbody>
</table>
| Task Description             | Chord transcription task requiring participants to annotate and segment the chord events in the MIREX09 chord transcription dataset. Please note that:  
  • Evaluations are performed at the triad level,  
  • results for both pretrained algorithms and algorithms trained and tested under 3 fold cross-validation are reported here.  
  • pretrained algorithms are likely to have been trained on the evaluation dataset hence they are expected to achieve higher results than algorithms evaluated on held out data. |
| Subject Metadata ID          | 26                                                                    |
| Subject Metadata Name        | Chord label sequence                                                 |
| Dataset ID                   | 33                                                                    |
| Dataset Name                 | MIREX09 Chord                                                        |
| Dataset Description          | MIREX2009 Chord transcription dataset composed of Christopher Hane’s Beatles dataset (C4DM, Queen Mary’s University of London) and Matthias Mauch’s Queen and Zweileck dataset (C4DM, Queen Mary’s University of London) |
| Date report generated        | Aug 6, 2010 7:56:08 PM                                               |

<table>
<thead>
<tr>
<th>Submission code</th>
<th>Submission name</th>
<th>Abstract PDF</th>
<th>Contributors</th>
</tr>
</thead>
<tbody>
<tr>
<td>CWB1</td>
<td>ChordID</td>
<td>PDF</td>
<td>Taemin Cho, Ron Weiss, Juan Bello</td>
</tr>
<tr>
<td>EW1</td>
<td>LabR05A Chord Train/Test 2010</td>
<td>PDF</td>
<td>Daniel Ellis, Adrian Weller</td>
</tr>
</tbody>
</table>
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

Audit
- Cassandra cluster volume store
- Solr index

Registry (WSO2)
- Algorithms
- Meandre Workflows
- Result Sets
- Collections

Compute resources

Storage resources
HTRC Architecture

**Portal Access**
- Blacklight

**Agent**
- HTRC Agent
  - Job Submission
  - Collection building

**Registry (WSO2)**
- Algorithms
- Meandre Workflows
- Result Sets
- Collections

**Solr**
- Proxy

**Compute resources**

**Storage resources**
- Cassandra cluster volume store
- Solr index

**Agent**
- Job Submission
- Collection building
HTRC Architecture

Portal Access
- Blacklight

Agent
- Job Submission
- Collection building

HTRC Registry

Registry (WSO2)
- Algorithms
- Meandre Workflows
- Result Sets
- Collections

Computes resources

Storage resources

HTRC Registry

Registry (WSO2)
- Algorithms
- Meandre Workflows
- Result Sets
- Collections

Solr index

Cluster volume store

WSO2 lean • enterprise • middleware
**HTRC Architecture**

**Portal Access**
- Blacklight

**Agent**
- Job Submission
- Collection building

**Security (OAuth2)**

**Registry (WSO2)**
- Algorithms
- Meandre Workflows
- Result Sets
- Collections

**Data**

**Compute resources**

**Secure Data API**

- RESTful Web Service
  - Language agnostic
  - Clients don’t have to deal with Cassandra
- Simple OAuth2 authentication
- HTTP over SSL
- Audits client access
- Protected behind firewall, accessible only to authorized IPs
The diagram illustrates the architecture of a system that includes several components:

- **Solr Proxy**
  - Solr proxy
  - Solr service
- **RFS distributed file system**
- **API access interface**
- **Storage resources**
  - Volume
  - Solr index
  - Solr service

The diagram also mentions direct programmatic access (by programs running on HTRC machines) and security (OAuth2). The system employs a Cassandra cluster volume store and a Solr index.

Additionally, the diagram shows connections to a Solr Proxy and a Solr service, indicating integration within a larger infrastructure.
Non-Consumptive Research-Secure Data Capsule
Meandre: Workbench Existing Flow

- Web-based UI
- Components and flows are retrieved from server
- Additional locations of components and flows can be added to server
- Create flow using a graphical drag and drop interface
- Change property values
- Execute the flow
Meandre Flow
Workset Creation for Scholarly Analysis (WCSA)
Workset Creation for Scholarly Analysis: Prototyping Project

• Collection analysis and prototype tools & services to facilitate work-set creation
  – J. Stephen Downie, Tim Cole, Beth Plale
  – Andrew W. Mellon Foundation
  – 1 July 2013 - 30 June 2015

• Proposal Narrative:
• The *Workset Creation for Scholarly Analysis: Prototyping Project* (WCSA) seeks to address three sets of tightly intertwined research questions regarding:

1. **Enriching** the metadata describing the HathiTrust corpus through mining of the resources themselves and leveraging end-user annotations;

2. **Augmenting** string-based metadata with URIs to leverage external services and Linked Open Data to facilitate discovery and the process of organizing HathiTrust resources into collections and worksets; and,

3. **Formalizing** the notion of collections and worksets in the context of the HathiTrust Research Center.
Motivation & Models

Collections, corpora, worksets, ...:

• Aggregations of items brought together in some context:
  – Archival
  – Curatorial
  – Experimental
  – Referential
  – Thematic (for research)

Carl Spitzweg. 1850
The Bookworm (Der Bücherwurm)
Grand Motivation

• The ability to slice through a massive corpus constructed from many different library collections, and out of that to construct the precise workset required for a particular scholarly investigation, is an example of the “game changing” potential of the HathiTrust...
What is a Workset?

1. A workset is an aggregation of materials brought together for the purpose of analysis.

2. Worksets are conceptual and must be expressible in a variety of ways
   - Need to allow creation outside of HathiTrust
   - Need to facilitate inclusion of resources beyond HathiTrust
   - Need to facilitate the inclusion of resources at many different levels of granularity beyond the book

3. Worksets encapsulate the specific materials that underwent analysis.
   - Need to capture provenance information
   - Possible recording of parameters

4. Worksets should be able to spawn descendants but otherwise immutable
Scope

HathiTrust Corpus

...Passages, Images, Annotations...

Pieces of Books

Metadata

Books

(Linked) Data / Metadata

Secondary Lit

Media

Workset (Research collection)

<serialized dataset(s)>

Scope
Dimensions of Workset Creation (Illustrative)

My work-set should contain (inspired by 2012 UnCamp):

- Volumes pertaining to Japan / in Japanese
- All volumes relevant to the study of Francis Bacon
- Music scores or notation extracted from HT volumes
- Volumes in HT similar to TCP-ECCO novels
- 19th c. English-language novels by female authors
- Representative sample (by pub date & genre) of French language items in HT
# MARC Metadata Shortcomings

<table>
<thead>
<tr>
<th>MARC Field</th>
<th>Percent of records in OCLC having instance of this field</th>
</tr>
</thead>
<tbody>
<tr>
<td>245 Title Statement</td>
<td>&gt; 99%</td>
</tr>
<tr>
<td>260 Publication Distribution, etc.</td>
<td>92%</td>
</tr>
<tr>
<td>500 General Note</td>
<td>41%</td>
</tr>
<tr>
<td>650 Topical Term / 653 Index Term – Uncontrolled</td>
<td>39% / 13%</td>
</tr>
<tr>
<td>050 LC Classification No / 082 Dewey Classification No</td>
<td>17% / 13%</td>
</tr>
<tr>
<td>655 Index Term -- Genre Form</td>
<td>12%</td>
</tr>
</tbody>
</table>

Table 2. Frequency of MARC fields in OCLC Records
### MARC Field Metadata Shortcomings II

<table>
<thead>
<tr>
<th>MARC Field</th>
<th>Percent of British Novel MARC records having instance of this field</th>
</tr>
</thead>
<tbody>
<tr>
<td>650 Topical Term</td>
<td>6%</td>
</tr>
<tr>
<td>050 LC Classification No / 082 Dewey Classification No</td>
<td>27% / 4%</td>
</tr>
<tr>
<td>655 Index Term -- Genre Form</td>
<td>5%</td>
</tr>
</tbody>
</table>

Table 3. Frequency of MARC fields used in 2,386 descriptions of 19th century British novels digitized from UIUC collections
Why Worksets?

- The result of a first-level, rough filter
- Better scale for intensive analytics
- Provides essential scope for certain analytics
  - Word frequency scope over Bacon’s essays
- Some tools (are trained to) work best on a narrow, homogeneous work-set
- Eliminate noise that would otherwise arise by asking questions across whole of HT
Research Questions (Illustrative only)

• Can we enrich the HathiTrust corpus metadata by distilling analytics over full text?

• Can we augment string-based metadata with URIs for recognized entities – e.g., names, subjects, publication location, etc. -- and by doing so can we leverage external services to facilitate discovery and clustering of resources?

• Can we leverage existing, well-defined external corpora to identify complementary subsets of HT volumes, and having done so can we demonstrate the ability to create and perform analytics over an integrated workset that includes resources external to HT?
Key Workset Questions

• Can we formalize the notion of collections and worksets in the HTRC context?

• What are the necessary elements of a “collection”? What are the necessary elements of a “workset”?

• How can we balance rigor with extensibility and flexibility?

• What roles do “data”, “metadata”, “annotations”, “tags”, “feature sets”, and so on, all play in the conception, creation, use and reuse of collections and worksets?
Two Project Streams

• Workset formal structures and semantics
  – Work in conjunction with Center for Informatics Research in Science and Scholarship at the Graduate School of Library and Information Science

• WCSA Prototyping Projects
  – Four projects funded by the grant but conducted by community teams
WCSA Timeline

- July 2013: Project Start
- Q1: User needs assessments / focus groups
- Q2: HT Corpus characterization Request For Prototype Proposals
- Q3: RFP Finalist Workshop (Chicago) February 20 Prototype experiment funding awarded
- Q4-6: Prototype experiments done Metadata workflow & work-set modeling
- Q7-8: Planning for prototype to production Report out
- June 2015: Project ends
Prototype Grants

As part of project, HTRC will make 4 sub-awards

• $40K awarded to each of 4 non-HTRC teams

• HTRC will collaborate with each team
  – Access to representative test data / metadata set
  – Collaborate on work with HT / HTRC APIs, etc.

<table>
<thead>
<tr>
<th>RFP &amp; Sub-Award Schedule</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013-11-15</td>
</tr>
<tr>
<td>2013-12-16</td>
</tr>
<tr>
<td>2014-01-15</td>
</tr>
<tr>
<td>2014-02-20</td>
</tr>
<tr>
<td>~ 2014-03-15</td>
</tr>
</tbody>
</table>
WCSA Summary

• Worksets are fundamental to the scholarly computational analysis enterprise

• We need a better understanding of their:
  – Constituent parts
  – Creation
  – Manipulation
  – Use and reuse

• Prototypes to lead to deeper tool development and metadata enhancement
Next Steps
Personal Goals for HTRC

• Engage in more collaborative projects
• Expand to have truly international partnerships
• Make sure to move beyond text
• Make sure to move beyond humanities!
Redux: Ongoing Challenges

How do we actually unlock the potential of 3 billion pages of human knowledge?

- Data quality issues
- Data structure challenges
- Metadata shortcomings
- Overcoming copyright barriers to research
  - Non-consumptive research
  - Computation to the data
- Moving beyond text
- Community building important and ongoing
Questions? Comments? Suggestions?

Special thanks to:
Jeremy York, Stacy Kowalczyk and Loretta Auvil