Building Open-Source Digital Curation Services & Repositories at Scale

Dr. Richard Marciano
Professor & Director of the Digital Curation Innovation Center (DCIC)

Greg Jansen / Will Thomas / Sohan Shah / Michael Kurtz
DCIC @ Maryland’s iSchool
University of Maryland

Presented on Feb. 20, 2018 at IDCC18
Session on Repository Services
1. Dvt. of Distributed Scalable NoSQL Catalogs and Repositories \textit{(DRAS-TIC)}

2. Dvt. of Cloud-Based Digital Curation Services \textit{(Brown Dog)}

3. Creation of a Testbed of Justice, Human Rights, and Cultural Heritage Collections

4. Devt. of a New Trans-Disciplinary Field: Computational Archival Science \textit{(CAS)}

5. Integration of Digital Curation Education & Research
1. Development of Distributed Scalable NoSQL Catalogs and Repositories (DRAS-TIC)

NoSQL distributed DB technology to support repositories that can scale out horizontally to 1000s of commodity servers. See: http://dcicblog.umd.edu/dras-tic-fedora/

Applies the new Fedora 5 API to the DCIC’s open-source software stack called DRAS-TIC. See: https://github.com/UMD-DRASTIC/drastic

Partners include:

1. Fedora Leadership Group and Steering Committee
3. University of Illinois Urbana-Champaign National Center for Supercomputing Applications (NCSA)
4. University of Maryland Libraries
5. Georgetown University Library
Digital Repository At Scale - That Invites Computation - To Improve Collections

- Product of 2-year startup by partners, Archival Analytics Ltd.
- Horizontal scaling to billions of files and beyond
- Web UI and command-line client
- Industry standard REST storage API (CDMI)
- Key-value metadata
- Eventing over MQTT message system
- Python source on GitHub (Open AGPL license)
- Based on the NoSQL Apache Cassandra (1,800 companies: CERN, eBay, GitHub, Hulu, Instagram, Netflix, Twitter, and scales to petabytes of storage and billions of objects)
2. Development of Cloud-Based Digital Curation Services (Brown Dog)

NSF “Brown Dog” Project, “The Super Mutt”
Public API for: (1) Format migration, and (2) Feature Extraction

Brown Dog: http://browndog.ncsa.illinois.edu/

Web-scale server virtualization
Part of an NSF DiBBs-funded project ($10.5M grant).

This service provides web and API access to 100s of tools, and is deployed over DRAS-TIC.

CNI Fall 2016 (slides and audio) – Computational Finding Aids:
https://www.cni.org/topics/digital-curation/drastic-measures-digital-repository-at-scale-that-invites-computation-to-improve-collections
3. Creation of a Testbed of Justice, Human Rights, and Cultural Heritage Collections

→ Accelerate the development of Digital Curation processes and services through the creation of a data observatory

Justice, Human Rights, & Cultural Heritage:

<table>
<thead>
<tr>
<th>THEME</th>
<th>PROJECTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Community Displacement</td>
<td>The Human Face of Big Data</td>
</tr>
<tr>
<td>Racial Zoning:</td>
<td>Mapping Inequality</td>
</tr>
<tr>
<td>Refugee Narratives:</td>
<td>St. Louis Voyage</td>
</tr>
<tr>
<td>Citizen Internment:</td>
<td>Japanese American WWII Camps</td>
</tr>
<tr>
<td>Movement of People:</td>
<td>Overseas Pension Project</td>
</tr>
<tr>
<td>Revealing Untold Stories:</td>
<td>Legacy of Slavery</td>
</tr>
</tbody>
</table>

Data Observatory:
- 100TB of data
- 100M files

NSF/NARA:
- 100TB
- 100M files
- 6,000 file types
- 150 fed agencies

Cyberinfrastructure for the Cur. & Mgt. of Dig. Assets at Scale:

<table>
<thead>
<tr>
<th>FUNDING</th>
<th>PROJECTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>NSF</td>
<td>Brown Dog</td>
</tr>
<tr>
<td>IMLS</td>
<td>DRAS-TIC Fedora</td>
</tr>
</tbody>
</table>
4. Devt. of a New Trans-Disciplinary Field: Computational Archival Science (CAS)

→ The Emergence of Computational XXX’s

- XXX=Social Science
  - “Investigating social and behavioral relationships and interactions through: social simulation, modeling, network analysis, and media analysis”, Wikipedia

- XXX=Biology
  - “The science of using biological data to develop algorithms or models to better understand biological systems”, Wikipedia

- XXX=Journalism
  - “Finding and telling news stories, WITH, BY, or ABOUT algorithms”, Nick Diakopoulos

- XXX=Archival Science ?
What is CAS?

A trans-disciplinary field concerned with the application of:

– computational methods and resources to large-scale records/archives:
  • processing, analysis, storage, long-term preservation, and access,
  • with the aim of improving efficiency, productivity and precision

– in support of appraisal, arrangement and description, preservation and access decisions, and engaging and undertaking research with archival materials.
An example (for Prof. Kohei Hatano!)


- Deep learning software to help botanists
- Botanical specimen categorization at museums (5 million specimens)
- Two big data analytics questions:
  1. With what accuracy can a trained neural network sort mercury-stained plant specimens from clean ones? [90-94%]
  2. With what accuracy can machine learning algorithms recognize members of two similar plant families? [96-99%]
“Computational Archival Science (CAS)” Portal
http://dcicblog.umd.edu/cas/

Goal: Explore computational treatments of archival and cultural content

CAS Founding Partners:
- United States:
  - UMD: Richard Marciano, Bill Underwood, Greg Jansen, Michael Kurtz
  - TACC: Maria Esteva
  - NARA: Mark Conrad
- Canada:
  - UBC: Vicki Lemieux
- United Kingdom:
  - KCL: Mark Hedges

Pursuing big ideas and new discoveries.
"Archival Records and Training in the Age of Big Data"

**Book:** “Advances in Librarianship – Re-Envisioning the MLIS: Perspectives on the Future of Library and Information Science Education”.

**Google Group:** computational-archival-science@googlegroups.com

1. Evolutionary prototyping and computational linguistics *(Bill Underwood)*
2. Graph analytics, DH and archival representation *(Richard Marciano)*
3. Computational finding aids *(Greg Jansen)*
4. Digital curation *(Michael Kurtz)*
5. Public engagement with (archival) content *(Mark Hedges)*
6. Authenticity *(Victoria Lemieux)*
7. Confluences between archival theory and computational methods *(Maria Esteva)*
8. Spatial and temporal analytics *(Mark Conrad)*
#3: Computational Curation of a Digitized Record Series of WWII Japanese-American Internment

- William Underwood, Richard Marciano, … — USA

**Computational Methods:** NLP, NER, GIS, Graph database

**Archival Concepts:** Digital curation, automated metadata extraction
GOAL:
Automating the review and release of records at scale
5. Integration of Digital Curation Education & Research

Key components of our initiative:

- Creating a new academic **Specialization, Archives & Digital Curation**, in the MLIS
- Organizing seminars for graduate students to define the theoretical and operational elements of **Computational Archival Science**
- Establishing a **Digital Curation for Information Professionals (DCIP) Certificate program**
- Offering students participation on **interdisciplinary digital curation projects**, at the intersection of archives, digital curation, Big Data, and analytics.

- E.g. Maryland State Archives’ **Legacy of Slavery** project
DCIC Digital Curation Projects…

“The active and ongoing management and enhancement of digital assets for current and future use.” Digital curation entails more than secure storage and preservation of digital information because curation may add value to digital information and increase its utility.

[Preparing the Workforce for Digital Curation (2015) - NRC / BRDI Report]

Archival Documents  Digitization  Datafabrication  Data Modeling  Data Visualization  Digital Storytelling

Archival Analytics
Interactive Tableau dashboard representing slavery statistics
Subset of the newspapers in which Runaway Slave ads were published. Pink nodes are newspapers, red nodes are “Slave” names, and purple nodes are “Owner” names. The relationships are “Appears_In”, and “Owned_By”.
Which slave names pertain to “Robert Bowie”. The purple node is the “Owner”, the red nodes are “Slave” names. The relationship is “Owned_By”.
CONTACTS

marciano@umd.edu

http://dcicblog.umd.edu/cas/
computational-archival-science@googlegroups.com