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

Digital Curation in the DCIC @ U. Maryland

1. Dvt. of Distributed Scalable NoSQL Catalogs and Repositories **(DRAS-TIC)**
2. Dvt. of Cloud-Based Digital Curation Services **(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 **(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

(2) Smithsonian Institution (Office of Research Info. Services and National Museum of American History)

(3) University of Illinois Urbana-Champaign National Center for Supercomputing Applications (NCSA)

(4) University of Maryland Libraries

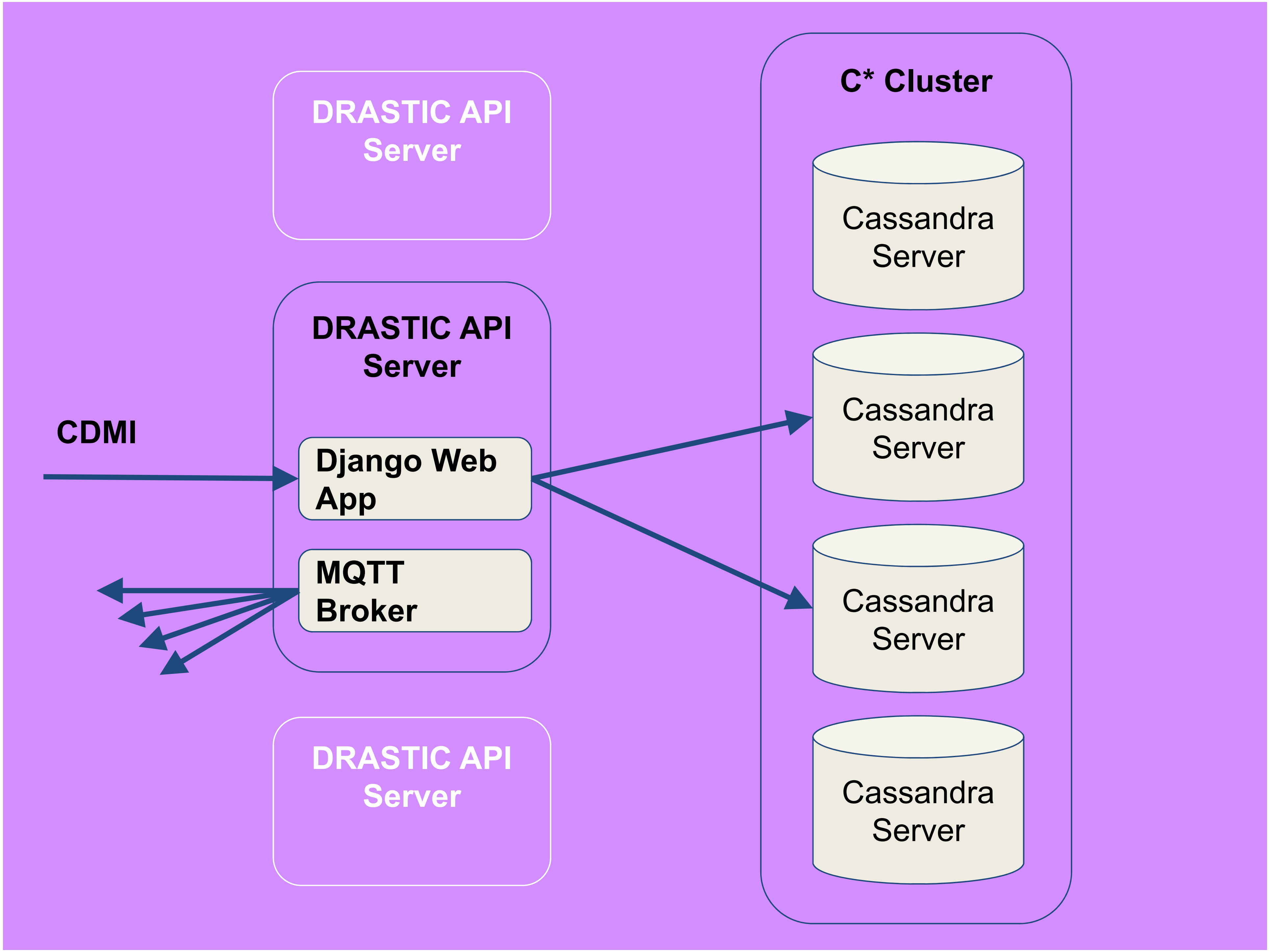
(5) Georgetown University Library

NEW

DRAS·TIC

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)*


DRASTIC API Server
C* Cluster
Cassandra Server
DRASTIC API Server
CDMI
Django Web App
Cassandra Server
MQTT Broker
Cassandra Server
DRASTIC API Server
Cassandra Server

2. Development of Cloud-Based Digital Curation Services (Brown Dog)

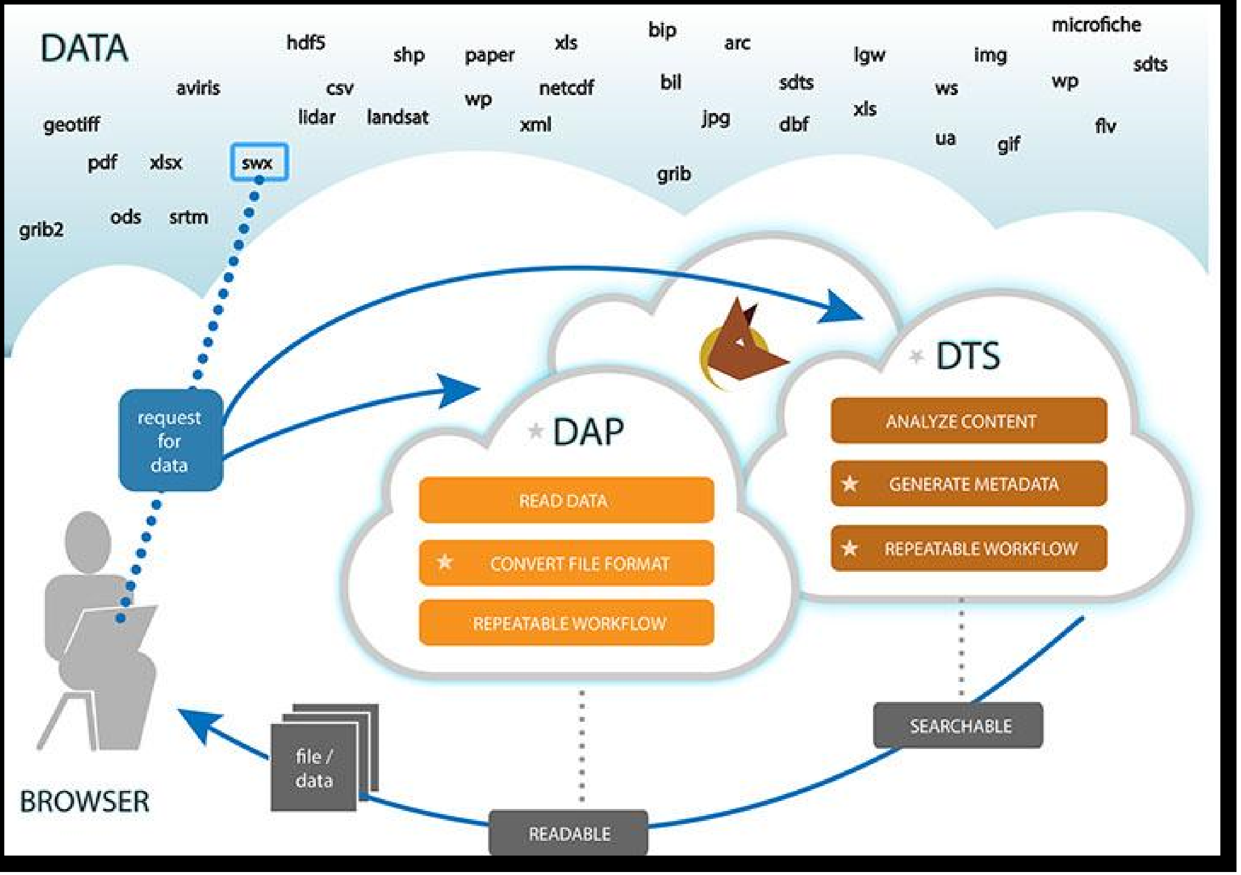
NSF "Brown Dog" Project, "The Super Mutt"
Public API for: (1) Format migration, and (2) Feature Extraction

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:

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

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:

FUNDING	PROJECTS
NSF	Brown Dog
IMLS	DRAS-TIC Fedora

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

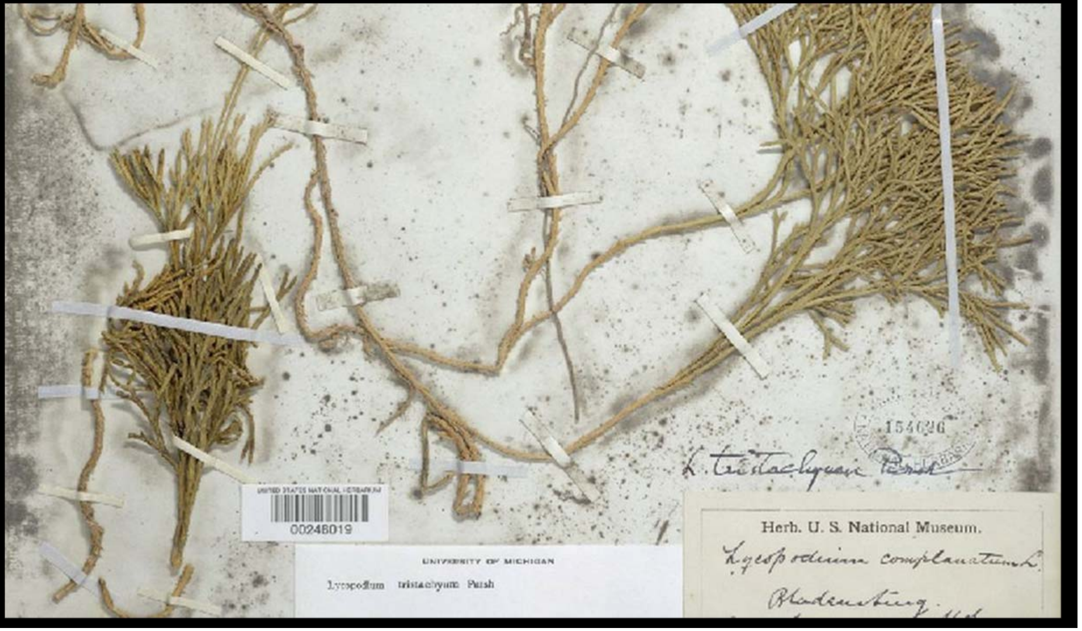
– in support of appraisal, arrangement and description, preservation and access decisions, and engaging and undertaking research with archival materials.

An example *(for Prof. Kohei Hatano !)*

How Artificial Intelligence Could Revolutionize Archival Museum Research *(Nov. 3, 2017)*

https://www.smithsonianmag.com/smithsonian-institution/how-artificial-intelligence-could-revolutionize-museum-research-180967065/

- 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%]


154626
L. tristachyum Pursh
UNITED STATES NATIONAL HERBARIUM
00248019
UNIVERSITY OF MICHIGAN
Lycopodium tristachyum Pursh
Herb. U. S. National Museum.
Lycopodium complanatum L.
Bladensburg.

“Computational Archival Science (CAS)” Portal

http://dcicblog.umd.edu/cas/

Goal: Explore computational treatments of archival and cultural content

Pursuing big ideas and new discoveries.

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

Foundational Book Chapter (June 2018)

"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)

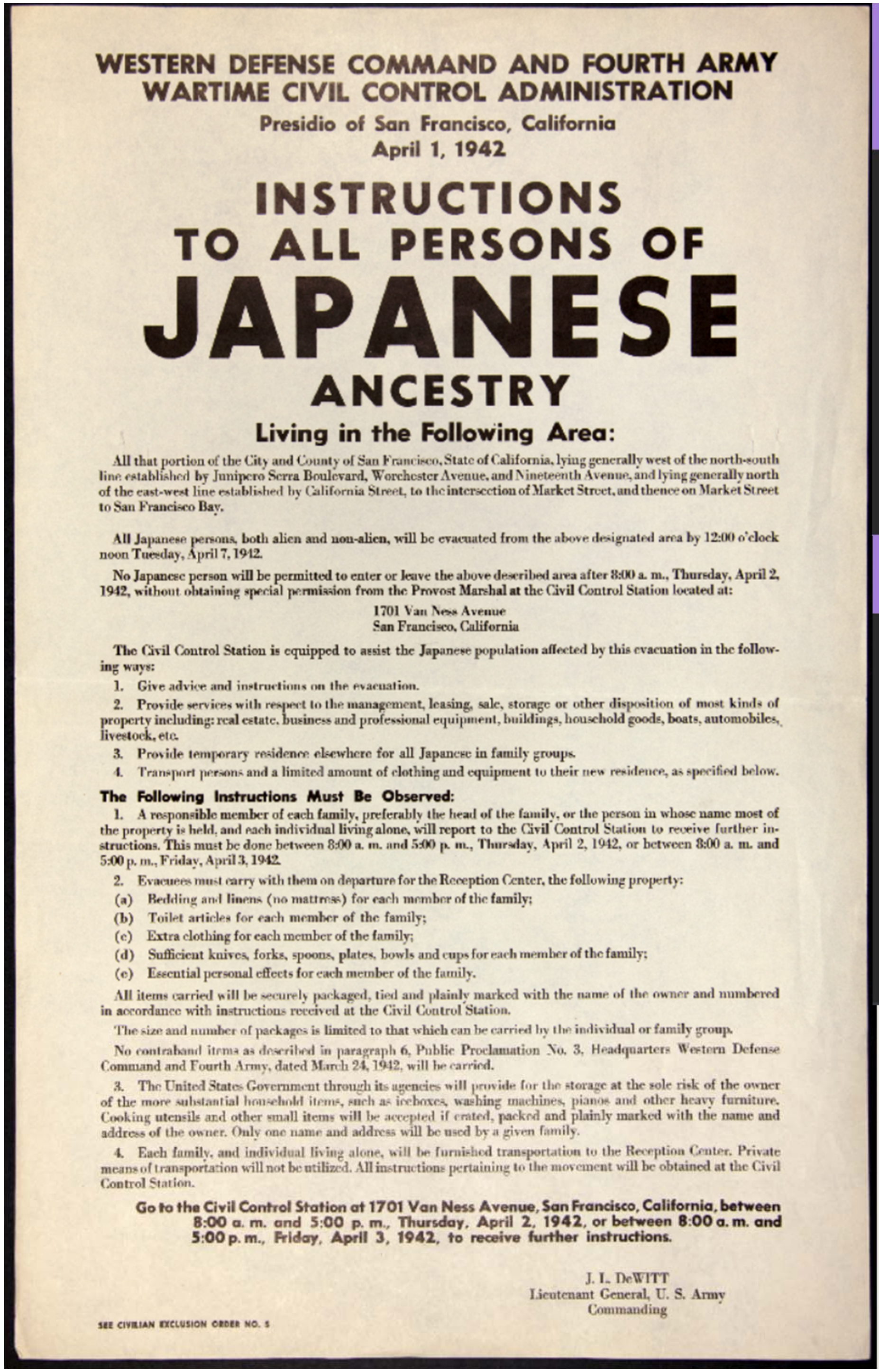
(8) Spatial and temporal analytics (Mark Conrad)

CAS 2017: http://dcicblog.umd.edu/cas/ieee_big_data_2017_cas-workshop/

- → #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



WESTERN DEFENSE COMMAND AND FOURTH ARMY
WARTIME CIVIL CONTROL ADMINISTRATION
Presidio of San Francisco, California
April 1, 1942

INSTRUCTIONS TO ALL PERSONS OF JAPANESE ANCESTRY

Living in the Following Area:

All that portion of the City and County of San Francisco, State of California, lying generally west of the north-south line established by Junipero Serra Boulevard, Worchester Avenue, and Nineteenth Avenue, and lying generally north of the east-west line established by California Street, to the intersection of Market Street, and thence on Market Street to San Francisco Bay.

All Japanese persons, both alien and non-alien, will be evacuated from the above designated area by 12:00 o'clock noon Tuesday, April 7, 1942.

No Japanese person will be permitted to enter or leave the above described area after 8:00 a. m., Thursday, April 2, 1942, without obtaining special permission from the Provost Marshal at the Civil Control Station located at:

1701 Van Ness Avenue
San Francisco, California

The Civil Control Station is equipped to assist the Japanese population affected by this evacuation in the following ways:

1. Give advice and instructions on the evacuation.

2. Provide services with respect to the management, leasing, sale, storage or other disposition of most kinds of property including: real estate, business and professional equipment, buildings, household goods, boats, automobiles, livestock, etc.

3. Provide temporary residence elsewhere for all Japanese in family groups.

4. Transport persons and a limited amount of clothing and equipment to their new residence, as specified below.

The Following Instructions Must Be Observed:

1. A responsible member of each family, preferably the head of the family, or the person in whose name most of the property is held, and each individual living alone, will report to the Civil Control Station to receive further instructions. This must be done between 8:00 a. m. and 5:00 p. m., Thursday, April 2, 1942, or between 8:00 a. m. and 5:00 p. m., Friday, April 3, 1942.

2. Evacuees must carry with them on departure for the Reception Center, the following property:

(a) Bedding and linens (no mattress) for each member of the family;

(b) Toilet articles for each member of the family;

(c) Extra clothing for each member of the family;

(d) Sufficient knives, forks, spoons, plates, bowls and cups for each member of the family;

(e) Essential personal effects for each member of the family.

All items carried will be securely packaged, tied and plainly marked with the name of the owner and numbered in accordance with instructions received at the Civil Control Station.

The size and number of packages is limited to that which can be carried by the individual or family group.

No contraband items as described in paragraph 6, Public Proclamation No. 3, Headquarters Western Defense Command and Fourth Army, dated March 24, 1942, will be carried.

3. The United States Government through its agencies will provide for the storage at the sole risk of the owner of the more substantial household items, such as iceboxes, washing machines, pianos and other heavy furniture. Cooking utensils and other small items will be accepted if crated, packed and plainly marked with the name and address of the owner. Only one name and address will be used by a given family.

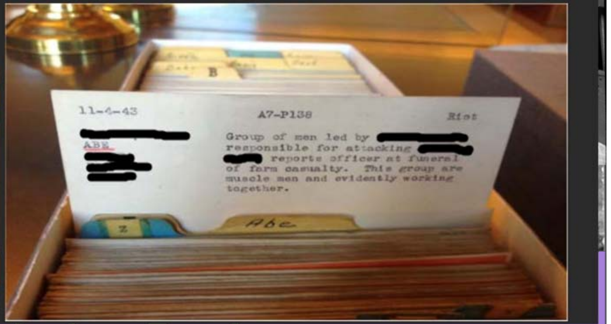
4. Each family, and individual living alone, will be furnished transportation to the Reception Center. Private means of transportation will not be utilized. All instructions pertaining to the movement will be obtained at the Civil Control Station.

Go to the Civil Control Station at 1701 Van Ness Avenue, San Francisco, California, between 8:00 a. m. and 5:00 p. m., Thursday, April 2, 1942, or between 8:00 a. m. and 5:00 p. m., Friday, April 3, 1942, to receive further instructions.

J. L. DeWITT
Lieutenant General, U. S. Army
Commanding

SEE CIVILIAN EXCLUSION ORDER NO. 5

Incident Index Cards

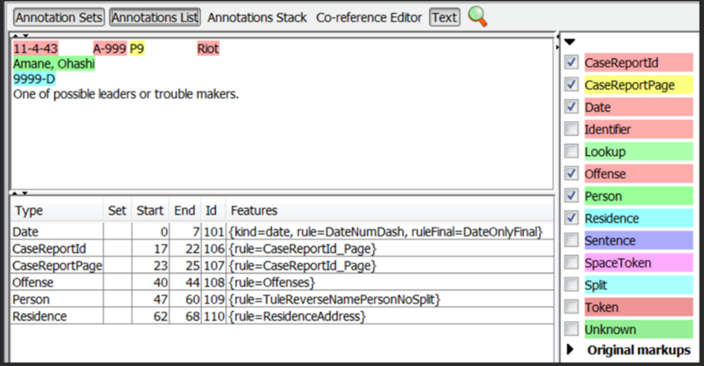


11-4-43 A7-P138 Riot

[redacted]
ABE
[redacted]

Group of men led by [redacted] responsible for attacking [redacted] [redacted] reports officer at funeral of farm casualty. This group are muscle men and evidently working together.

List for a Sample Index Card

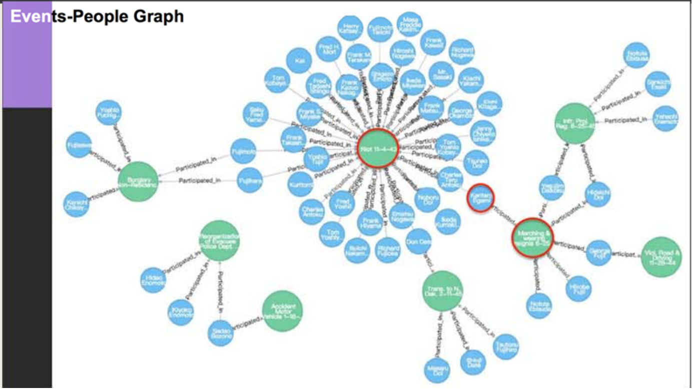


11-4-43 A-999 P9 Riot
Amane, Ohashi
9999-D
One of possible leaders or trouble makers.

Type	Set	Start	End	Id	Features
Date		0	7	101	{kind=date, rule=DateNumDash, ruleFinal=DateOnlyFinal}
CaseReportId		17	22	106	{rule=CaseReportId_Page}
CaseReportPage		23	25	107	{rule=CaseReportId_Page}
Offense		40	44	108	{rule=Offenses}
Person		47	60	109	{rule=TuleReverseNamePersonNoSplit}
Residence		62	68	110	{rule=ResidenceAddress}

☑ CaseReportId
☑ CaseReportPage
☑ Date
☐ Identifier
☐ Lookup
☑ Offense
☑ Person
☑ Residence
☐ Sentence
☐ SpaceToken
☐ Split
☐ Token
☐ Unknown
▶ Original markups

GOAL:
Automating the review and release of records at scale



Events-People Graph

5. Integration of Digital Curation Education & Research

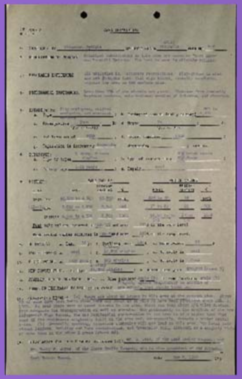
Key components of our initiative:

- 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

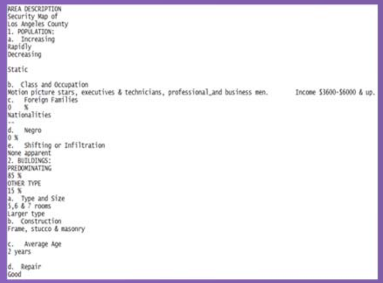
DCIC Digital Curation Projects…

"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



AREA DESCRIPTION
Security Map of
Los Angeles County
1. POPULATION:
a. Increasing
Rapidly
Decreasing

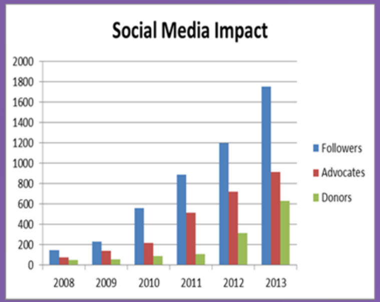
Static

b. Class and Occupation
Motion picture stars, executives & technicians, professional_and business men. Income $3600-$6000 & up.
c. Foreign Families
0 %
Nationalities
--
d. Negro
0 %
e. Shifting or Infiltration
None apparent
2. BUILDINGS:
PREDOMINATING
85 %
OTHER TYPE
15 %
a. Type and Size
5,6 & 7 rooms
Larger type
b. Construction
Frame, stucco & masonry

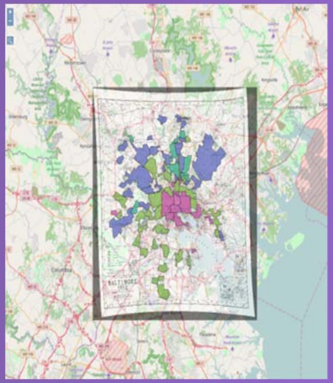
c. Average Age
2 years

d. Repair
Good

Digitization



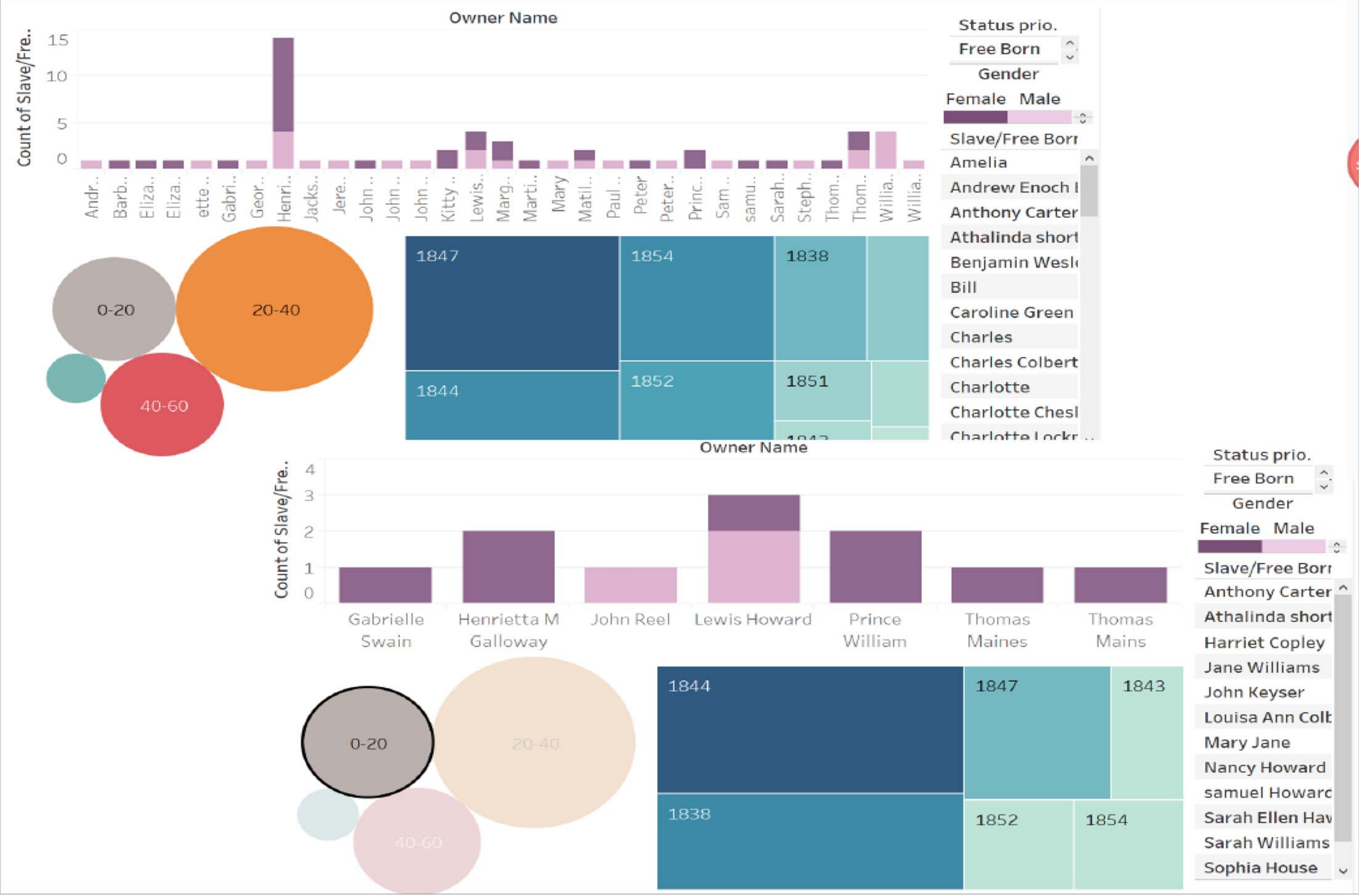
Datafication
Data Modeling



Data Visualization
Digital Storytelling

Archival Analytics

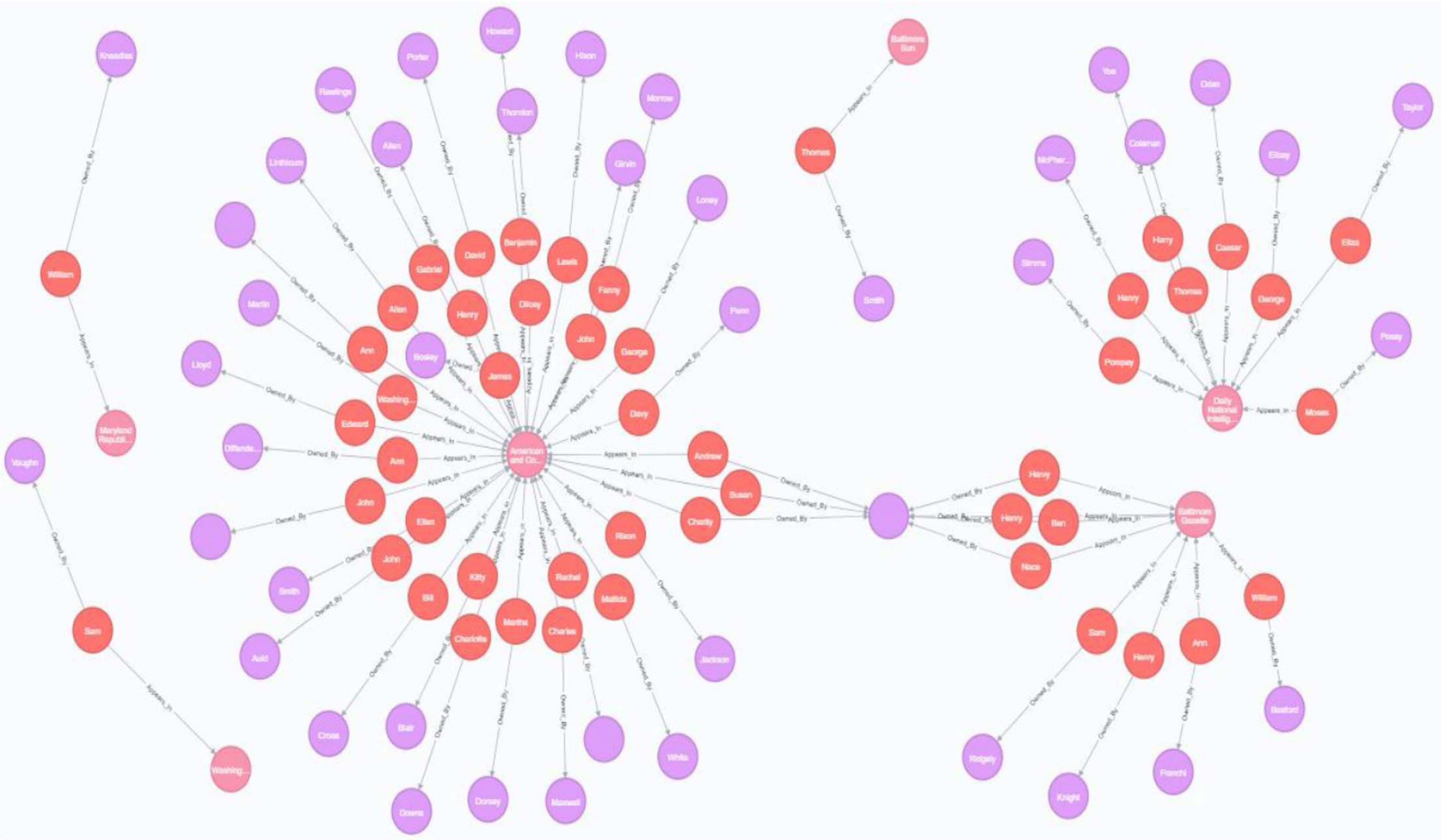
Interactive Tableau dashboard representing slavery statistics



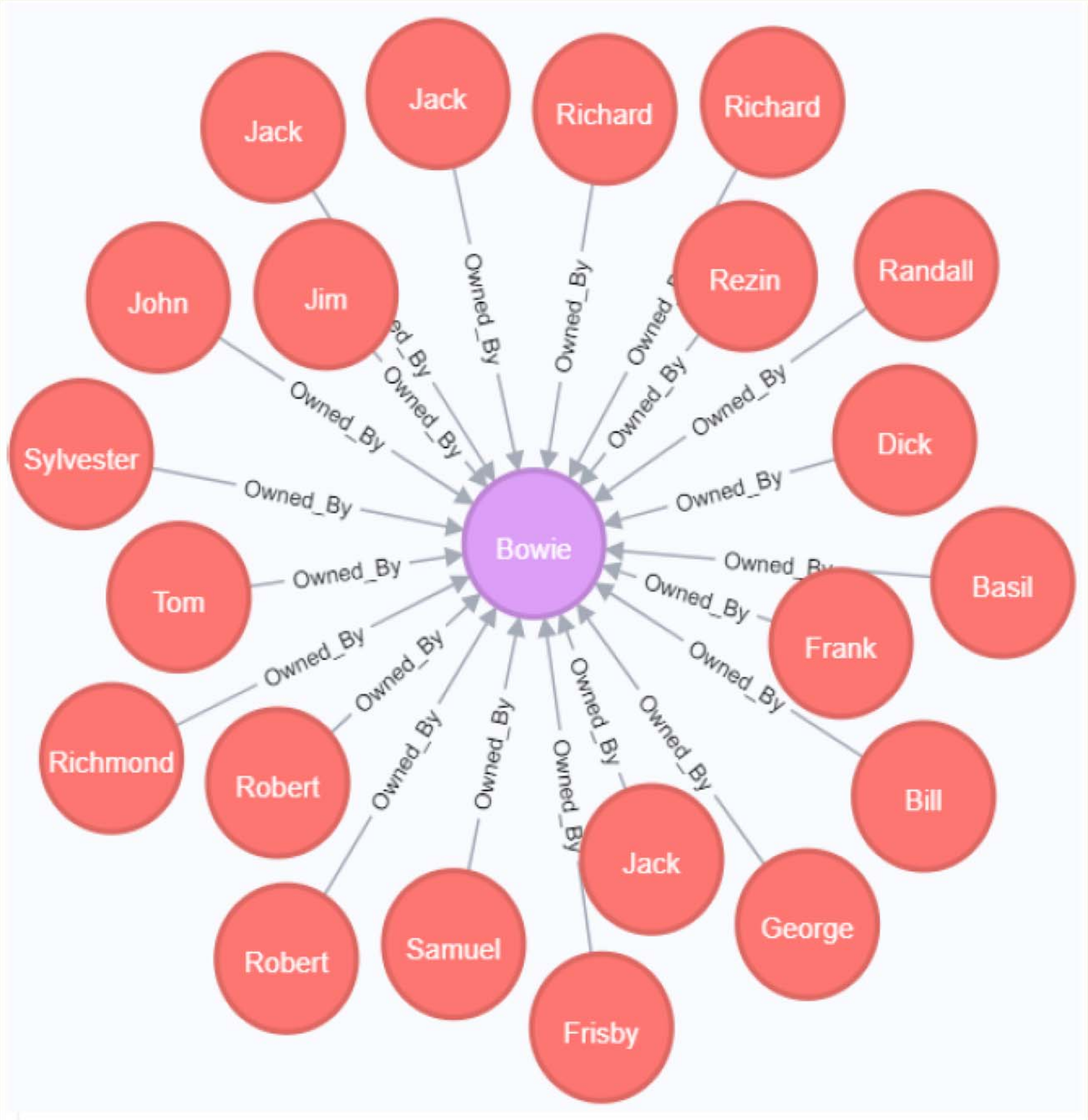
Status prio.
Free Born

Status prio.
Free Born

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