Practical Digital Curation Skills for Archivists in the 21st Century

Myeong LEE
Mary KENDIG
Richard MARCIANO
Greg JANSEN

U. Maryland iSchool DCIC

MARAC FALL 2016

ANNAPOLIS, MD | NOVEMBER 3-5, 2016
Agenda

○ Introduction to DCIC
○ Defining Digital Curation
○ Case Studies
  1. Human Face of Big Data [Community Displacement]
  2. Mapping Inequality [Racial Zoning]
  3. St. Louis Voyage [Refugee Narratives]
  4. WWII Japanese American Camps [Citizen Internment]
○ Computational Archival Science Themes
○ Practical Skills for 21st Century Archivists
○ Conclusion / Questions
ARC: Archives Research & Collaboration Lab

Director: Ricky Punzalan

ARC studies and develops innovative approaches, systems, strategies, and tools to foster sustainable futures for archives, preservation, and digital curation.

http://archivescollaboratory.umd.edu/

SALT: Sustainable Archives & Leveraging Technologies

Director: Richard Marciano

SALT is an interdisciplinary lab, which focuses on the long-term preservation of digital cultural and research assets at scale. SALT is an acronym for Sustainable Access and Longevity of Digital Information and uses as its logo the two thousand year-old ancient Chinese pictograph for salt (“yan”) which is a metaphor for the integration of policy, governance, infrastructure, and content.

http://salt.umd.edu

curate lab
Hornbake South 4110
Digital lab for group learning, collaborative design, and hands-on digital curation project development (23 seats, 3 interactive screens, 12 workstations with 12TB of storage).

digitization lab
Hornbake South 4110D
Document scanning, image manipulation, and archival inventory facility for group projects.

server farm
UMD Computer & Space Sci. Bldg
On-campus virtual machine farm for research data processing, storage, and hosting (15TB storage, 2 Dell servers, VMware-powered).

cloud lab
Amazon Cloud
Dashboard-enabled virtual computing lab in the cloud for creating Windows/Linux instances using Amazon Web Services (AWS).

data Cave
UMD Cyberinfrastructure Center at the RiverTech Bldg
Digital Repository AI Scale That Invites Computation (To Improve Collections): a petabyte scale archival storage and preservation repository (based on DRAS-TIC: open-source software (NoSQL, Cassandra database) and computational infrastructure (4 Dell nodes)).

dcic digital curation innovation center

http://dcic.umd.edu

Mission

Be a leader in the digital curation research and educational fields, and foster interdisciplinary collaborations using Big Records and archival analytics with public/industry/government partnerships.

Sponsor interdisciplinary projects that explore the integration of archival research data, user-contributed data, and technology to generate new forms of analysis and historical research engagements, particularly in the arenas of social justice, human rights, and cultural heritage.
Projects

Cyberinfrastructure for the curation & management of digital assets at scale:

"Brown Dog"
A CIC Big $10.5M NSF/DIBBS-funded collaboration with U of Illinois NCSA Supercomputing Center and industry partners (NetApp and Archive Analytics Solutions). This project aims to help accelerate the development of digital curation processes and services and create a data observatory to provide access to Big Records training sets and teach students practical digital curation skills.

"Curate Cloud"
A $300K IMLS-funded project that helped launch a new online professional education certificate for digital curation professionals, the Curation and Management of Digital Assets (CMDA). Curate Cloud is also developing an open-source research and educational platform, the VCI (Virtual Computing Lab), to remove barriers to access for curation tools and resources.

Digital Curation training:

Digital Curation Fellowships
The School has several Fellowship opportunities for students in digital curation and archives. These include a collaboration with the National Agricultural Library (NAL); extensive project work with the National Park Service (NPS); and a scholarship established in honor of Bruce Ambach, retired senior archivist and School faculty member.

Interdisciplinary Research Teams
Gain new digital skills, conduct interdisciplinary research, explore professional development opportunities at the intersection of archives, big data, and analytics through a number of project themes: Refugee Narratives, Community Displacement, Racial Zoning, Cyberinfrastructure for Digital Curation, Movement of People, Citizen Internment.

People

Research Staff:
Richard Marciano, Director & SALT Lab Director
Michael Kurz, Associate Director
Ricardo Ponzanelli, Research Associate & ARC Lab Director
Ken Neiger, Research Associate & Digital Lab Director
Greg Jensen, Research Software Architect
Maria Elenova, Affiliate Professor
Victoria Lemieux, Affiliate Professor
William Rinderbush, Affiliate Professor

Mary Kendig / Myeong Lee, Graduate Research Assistants

Research Affiliates:
U. Maryland
Tammy Clegg, Nick Diakopoulos, Jesse Johnston, Trevor Owens, Jenny Preece, Katio Shilton
External
Bruce Ambach, Natalie Baur, John Barns, Andrew Law, Scott Madry

Postdoctoral Fellows:
Morgan Dansell, Kathryn Guccia, Adam Kriesberg (Advisor: Ponzanelli)

Students (undergraduate, master’s [MLIS, MRM, JAN], doctoral):

Doctoral Students:
Andrew Casperino, William Thomas
Diana Obzva (Advisor: Marciano)
Edward Summers, Amy Wicken

Projects

Justice, Human Rights, & Cultural Heritage:

Overseas Pension Project
A student- and professional society-sponsored project to collect information documenting payment of pensions to American veterans living overseas. The project creates datasets documenting migration patterns, the flow of money, health conditions, and family connections prior to World War I.

International Research Portal Project (IRP)
This project will improve access to an important tool which identifies and locates looted art and other cultural assets found on the International Research Portal for Records Related to Nazi-Era Cultural Property.

Mapping the Voyage of the St. Louis
In 1939, 937 passengers (mostly Jews) fled Germany aboard the SS S.S. Louis ship, heading to Cuba, where they were turned away and forced to return to Europe where 254 were killed during the Holocaust. The project looks at mapping individual and collective stories through graph database techniques.

Japanese-American WWII Camps
Building on a UMD RIA Seed Grant, the project explores the integration of archival and user-contributed data using social networking graphs to link people, places, and events. Using WWII Camp data.

Mapping Inequality
A project with Johns Hopkins, Virginia Tech, and U. of Richmond where a national collection of New Deal rehiring data is being crowdsourced (these unique records capture racial, ethnic, and economic conditions).

The Human Face of Big Data
A student-led project that will create access and collaborative opportunities around historically and socially significant heterogeneous datasets rooted in urban renewal housing records for a number of cities.
Mission:

• Be a leader in the digital curation research and educational fields, and foster interdisciplinary partnerships using **Big Records and archival analytics** through public / industry / government collaborations.

• Sponsor interdisciplinary projects that explore the integration of archival research data, user-contributed data, and technology to generate new forms of analysis and historical research engagements.
"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]
1. Human Face of Big Data

[Community Displacement]

UMD Student Team:
Myeong Lee
Shiyun Chen
David Zhang
Edel Spencer
Rajesh Gnanasekaran
Alicia Geller
Hardik Jhaveri
Urban Renewal Documents

How each property was acquired by Housing Authority?

How did the old neighborhood look like?

Providing a forest of how the urban renewal project was conducted in a city through building a system.
Digitization Process of an Old Map

Paper Map → Scanning & Adjusting → Georeferencing → Geo-tracing
Scanning and Adjusting

Distortions Due to Tiled Scanning

Smart Object in Photoshop
Georeferencing

Using **ArcGIS** or **QGIS**

**Aligning** an Old Raster Map onto the Modern Online Map

Used another clean Map **as a Guideline** to Align the Scanned Map
Geo-Tracing

The process of creating computer-detectable polygons on top of the raster map.

Possible using QGIS, ArcGIS, or Leaflet.draw.

Stored in a Shapefile.

Shapefile can be exported to GeoJSON.
Iterative System Design and Prototyping

User Persona

User Scenario

Data Modeling

User Feedback

Interface/DB Design (Wired-frame)
## Human Face of Big Data

<table>
<thead>
<tr>
<th>About</th>
<th>Human Face Map</th>
<th>Resources</th>
<th>Data</th>
<th>Contact us</th>
</tr>
</thead>
</table>

![Map Image]

**Data**

- **Year**: 1965, 1978
- **Name:**
- **Address:**
- **Search**
- **Advanced Search**
- **Dynamic Area:**
  - **Statistics**
  - **Property Information List:** Name, Address
2. Mapping Inequality
[Racial Zoning]

UMD Student Team:
Mary Kendig
Myeong Lee
Sydney Vaile
Maddie Allen
Martin Almirón
Jhon De La Cruz
Shaina Destine
Erin Durham
Darlene Reyes
Benjamin Sagay
Richard Bool
Historical Context

- Home Owners Loan Corporation 1930’s - 1940’s
  - Rated neighborhoods by racial makeup
  - Areas without loans fell apart
- 1950’s Urban Renewal targeted areas for clearance
- Result: Mass displacement
- RG195: Federal Home Loan Bank Board, HOLC, 1933 - 1951
- Contains Maps, Neighborhood Surveys, Loan Information
Mapping Inequality

Documents
- Each survey corresponded to city map
  - **Green**: White / Wealthy = Best
  - **Blue**: White / Working = Still Desirable
  - **Yellow**: Foreign / Increase in PoC = Declining
  - **Red**: Black and Hispanic = Hazardous

Collection Statistics
- 150 Boxes
- Over 6,000 surveys alone
- 250 cities
AREA DESCRIPTION

Security Map of LOS ANGELES COUNTY

1. POPULATION:  a. Increasing  
    b. Class and Occupation: Jewish professional & business men, Mexican workers, WPA workers, etc. Income $700 to $2000 and up
    c. Foreign Families 50% Nationalities: Russian, Polish & Armenian Jews, Slavs, Greeks, American Mexicans, Japanese and Italians
    d. Negro 1% Subversive racial elements increasing.
    e. Shifting or Infiltration

8. DESCRIPTION AND CHARACTERISTICS OF AREA:
   Terrain: Level to hillside with generally favorable grades and comparatively few construction hazards. Land improved 90%. This is a "melting pot" area and is literally honeycombed with diverse and subversive racial elements. It is seriously doubted whether there is a single block in the area which does not contain detrimental racial elements, and there are very few districts which are not hopelessly heterogeneous in type of improvement and quality of maintenance. Schools, churches, trading centers, recreational areas and transportation are all conveniently available. Many of the thoroughfares are arterial in character and constitute traffic hazards. This area is wholly in the City of Los Angeles. It is hazardous residential territory and is accorded a general medial red grade, although in many parts slum conditions prevail. The Federal Government, in conjunction with the city government are undertaking a slum clearance project covering 41 areas in the extreme northeast part of the area.

AREA DESCRIPTION

1. POPULATION
   a. Ethnic Origin
   b. Density
   c. Age Structure
   d. Income Levels

2. BUILDINGS
   a. Type and Size
   b. Construction
   c. Average Age
   d. Value

3. NEW CONSTRUCTION
   a. Type of Property
   b. Age

4. SALE OF HOME PROPERTIES
   a. Median Sale Price
   b. Average Sale Price

5. MORTGAGE FUNDS
   a. Average Mortgage Amount

6. DESCRIPTION AND CHARACTERISTICS OF AREA
   a. Land Use
   b. Demographics
   c. Infrastructure

LOCATION

[Map and data tables with various demographic and economic indicators]
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.
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

Python Programming Script
import csv
import re
import collections
import json

txt_file = "r"/Users/myeong/git/ICIC/0330_Python_Session/a-001.txt"
out_file = "r"/Users/myeong/git/ICIC/0330_Python_Session/a-001.json"
out_csv = "r"/Users/myeong/git/ICIC/0330_Python_Session/a-001.csv"

# file_content = open(txt_file).read()
# tokens = nltk.word_tokenize(file_content)
parsed_data = {}
section = 0

with open(txt_file) as f:
    for line in f:
        line = line.strip()
        line = ''.join(line.split())
        if line.startswith("Security Map of "):
            section = 1
            parsed_data["0.0"] = ''.join(f.next().strip().split())
        elif line.startswith("a. Increasing"):  
            parsed_data["1.a.1"] = 
            join(f.next().strip().split())
        elif line.startswith("Decreasing"):  
            parsed_data["1.a.2"] = 
            join(f.next().strip().split())
        elif line.startswith("Static"):  
            parsed_data["1.a.3"] = 
            join(f.next().strip().split())
        elif line.startswith("b. Class and Occupation"):  
            parsed_data["2.b"] = 
            join(f.next().strip().split())

"
```python
import csv
import re
import collections
import json

txt_file = r"/Users/myeong/git/ICIC/0330_Python_Session/a-001.txt"
out_file = r"/Users/myeong/git/ICIC/0330_Python_Session/a-001.xlsx"
out_csv = r"/Users/myeong/git/ICIC/0330_Python_Session/a-001.csv"

# file_content = open(txt_file).read()
# tokens = nltk.word_tokenize(file_content)

parsed_data = {}
session = 0

with open(txt_file) as f:
    for line in f:
        line = line.strip()
        line = ' '.join(line.split())
        if line.startswith("Security Map of ":
            section = 1
            parsed_data["0.0"] = ' '.join(f.next().strip().split())
        elif line.startswith("A. Increasing":)
            parsed_data["1.0.1"] = ' '.join(f.next().strip().split())
        elif line.startswith("A. Decreasing":)
            parsed_data["1.0.2"] = ' '.join(f.next().strip().split())
        elif line.startswith("B. Class and Occupation":)
            parsed_data["1.0.3"] = ' '.join(f.next().strip().split())
```

---

**Excel Spreadsheet**

<table>
<thead>
<tr>
<th>Contract</th>
<th>Task Order</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>FY12 Computing Services</td>
<td>CSD</td>
<td>Information Assurance</td>
</tr>
<tr>
<td>NS28012012</td>
<td>HIC082010F721</td>
<td>NAVFAC Custodial &amp; Repairs</td>
</tr>
<tr>
<td>NAVFAC Support Agreement</td>
<td>PMA11005</td>
<td>Pensacola Support Agency</td>
</tr>
<tr>
<td>Support Agreement</td>
<td>PMA_11_001</td>
<td>Pensacola Support Agency</td>
</tr>
<tr>
<td>‘or L’</td>
<td>0000</td>
<td></td>
</tr>
<tr>
<td>0000</td>
<td>0001</td>
<td>This is a test entry for the test</td>
</tr>
<tr>
<td>027017</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>111111110000</td>
<td>0001</td>
<td>Test Contract</td>
</tr>
<tr>
<td>12345677</td>
<td>0000</td>
<td></td>
</tr>
<tr>
<td>17717</td>
<td>0001</td>
<td>TEST for ORG</td>
</tr>
<tr>
<td>1C105426D0X</td>
<td>0000</td>
<td></td>
</tr>
<tr>
<td>20050821362</td>
<td>0000</td>
<td></td>
</tr>
<tr>
<td>20070202601</td>
<td>0000</td>
<td></td>
</tr>
<tr>
<td>2603201253</td>
<td>0000</td>
<td>Dell PowerEdge Server</td>
</tr>
<tr>
<td>704054240A</td>
<td>MULTIPLE</td>
<td></td>
</tr>
<tr>
<td>710454265A</td>
<td>MULTIPLE</td>
<td></td>
</tr>
<tr>
<td>1711954256</td>
<td>0000</td>
<td></td>
</tr>
<tr>
<td>Adobe 2013</td>
<td>0</td>
<td>Adobe licenses for eLearning</td>
</tr>
<tr>
<td>APRIL Support Agreement</td>
<td>PMA12021</td>
<td>APRIL Annual ISW Maint</td>
</tr>
</tbody>
</table>

Mapping Inequality US

Link: https://dsl.richmond.edu/panorama/redlining/

Partner Universities:
U. Maryland’s iSchool’s DCIC Center
Dr. Richard Marciano

U. of Richmond's Digital Scholarship Lab
Dr. Robert Nelson

Virginia Tech
Dr. LaDale Winling

Johns Hopkins University
Dr. N.D.B. Connolly
SCRIBE: Crowdsourcing Platform
SCRIBE: Crowdsourcing Platform

Originally Developed by NYPL (https://github.com/zooniverse/scribeAPI)

Open Source Software written in Ruby on Rails

Collaborative Transcription and Verification of Texts that cannot be read automatically: CrowdSourcing

Automatic Generation of Marking Boxes
3. St. Louis Voyage
[Refugee Narratives]

UMD Student Team:
Mary Kendig
Myuresh Amdekar
Brian Refford
Sohan Shah
Pal Doshi
Yuting Liao
Ruchira Kapoor
Sohan Shah
3. St. Louis Voyage [Refugee Narratives]

- May 13, 1939 German ship SS St. Louis sailed 937 Jewish refugees from Third Reich to Havana, Cuba
- Despite arriving, SS St. Louis turned away
- Captain sailed to US
- State Department and White House denied entry
St. Louis Voyage

- United States Holocaust Memorial Museum determined fate of all 937 passengers
- Developed spreadsheets containing passenger data, events, and countries
- Continuously collecting photographs, stories, and documents
St. Louis Voyage: Building Archival Databases

- Students working directly with USHMM staff to conduct requirements gathering
- Using SQL programming and MySQL Workbench
- Building application for users to interact with data and collections
St. Louis Voyage: Data Analytics

- Using R Programming to conduct data analytics
- Working with visualization software Tableau, NodeXL, Gephi
- Attributing coordinates to camps and cities to visualize movement of families
Network of People in Internment

**Event Type:** Interment  **Country:** France

By Yuting Lao

- 1939–40 (before Nazi invasion)

1940-46 (after Nazi invasion)
4. WWII Japanese-American Camps
[Citizen Internment]

UMD Student Team:
Jiahui Wu
Kelsey Diemand
Meaghan Wilson
Vinita Atre
Diane Travis

UMD Faculty:
Richard Marciano
Bill Underwood
AETN's "In Their Words" project re: the Rohwer Internment Camp:

- Four-year old George Takei and his family were moved from CA to the Rohwer, AK, Relocation Center in 1942. The family members lived in Arkansas about eight months. Following the refusal of Takei's parents to swear loyalty to the United States, the family was moved to the maximum-security camp at Tule Lake, CA, where they stayed until 1946.

NPR / On the Media:

- Talks with great clarity about the FDR executive order and the loyalty oath...
Assembly Centers, Relocation Centers, Justice Department Internment Camps, & Citizen Isolation Camps
Background:
Tule Lake Segregation Center
Welcome to the Tule Lake Unit of WWII Valor in the Pacific National Monument

The Tule Lake Unit, WWII Valor in the Pacific National Monument includes both the Tule Lake Segregation Center, the largest and most controversial of the sites where Japanese Americans were incarcerated during World War II, and Camp Tulelake, which was first a Civilian Conservation Corps camp, then a prisoner of war camp, and finally an additional facility to detain Japanese Americans.
Incident Cards

KAZAMA, Tonia
ABE
OSHITA
YAMAMOTO
OSHITA

11-4-43
A7-P138
Riot

Group of men led by Tonia Kazama responsible for attacking John D. Cook, reports officer at funeral of farm casualty. This group are muscle men and evidently working together.
Digital Curation Workflow for Incident Cards by NARA

- Scan and OCR the "incident cards".
- Extract the names from the OCR text and capture in a database.
- Check the names against the full database of internees (also part of RG210), which includes birthdates.
- Identify names as being those of juveniles (under 18) and redact them from the database of names and from the scanned card images before their delivery to the iSchool DCIC Center.

Acknowledging the 2015 support of NARA’s:

- Office of Innovation: Pam Wright, Markus Most, John Martinez, Denise Henderson
- Legislative Archives & Museum Services / Exhibits Dept.: Michelle Farnsworth
- Federal Records Center, St. Louis: Richard Morgan, Bob Marsh
- Research Services: Chris Naylor & Martha Murphy [FOIA b(6) review]
Digital Curation Processing Workflow

**Cultural Object Digitization**
- NARA Scan Lab
- DCIC digitization lab

**Optical Character Recognition (OCR)**
- KoFax Express
- Tesseract
- Cuneiform Linux
- ABBYY FineReader

**Name Entity Recognition (NER)**
- Alchemy API
- OpenNLP
- Stanford NER
- OpenCalais

**Database**
Phase II

- **Developing tools for optimized scalable input and processing** for this and other NARA collections

- **Automatic content extraction** (recognition of named entities, case ids and events) -- using GATE software

- **Structuring information** from notecards (XML Note Card document type)

- **Integration of records through analytics and visualization**
11-4-43  A-7 Page 842  Riot

Abe, Tetsuo  F. N. 1578-A

On 3-16-44 the Coordinating Committee recommended that this man be released from the stockade. They had no information to submit.
Incident Card Modeling

- Date
- Type
- Detail
- Name
- Political group
- Id
- Type
- Money involved
- Subsequent event
- Location
- Take part in
- Reside
- Camp

(Japanese inmate/official/job title)
Analyzing the Panama Papers with Graph Databases:
[https://neo4j.com/blog/analyzing-panama-papers-neo4j/](https://neo4j.com/blog/analyzing-panama-papers-neo4j/)

1. Acquire docs
2. Classify docs
   - Scan / OCR
   - Extract metadata
3. Data modeling
4. NER extraction
5. Relationship building
6. Graph analytics

**The structure of the leak**
The 11.5 million contain the following file types:

<table>
<thead>
<tr>
<th>File Type</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emails</td>
<td>1,359,968</td>
</tr>
<tr>
<td>Database files</td>
<td>2,131,765</td>
</tr>
<tr>
<td>PDFs</td>
<td>1,117,896</td>
</tr>
<tr>
<td>Images</td>
<td>1,242</td>
</tr>
<tr>
<td>Text documents</td>
<td>325,196</td>
</tr>
<tr>
<td>Other</td>
<td>2,242</td>
</tr>
</tbody>
</table>

**Family of the Azerbaijan’s President**
Digital Curation
PROJECT
OBJECTIVES:

(1) Gain new digital skills

(2) Conduct interdisciplinary research

... at the intersection of archives, big data, & analytics
# How Each Project Is Related to Computational Archival Science (CAS) Themes:

<table>
<thead>
<tr>
<th>Project</th>
<th>Computational Linguistics</th>
<th>Data Modeling &amp; Evolutionary Prototyping</th>
<th>Graph Analytics</th>
<th>Crowdsourcing</th>
<th>GIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Human Face of Big Data [Community Displacement]</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>
Workshop Title: Computational Archival Science: digital records in the age of big data

Venue: Thursday, Dec. 8, 2016 @ Hyatt Regency Washington on Capital Hill

- **Analytics in support of archival processing**, including appraisal, arrangement and description.
- **Scalable services for archives**, including identification, preservation, metadata generation, integrity checking, normalization, reconciliation, linked data, entity extraction, anonymization and reduction.
- **New forms of archives**, including Web, social media, audiovisual archives, and blockchain.
- **Cyber-infrastructures for archive-based research** and for development and hosting of collections.
- **Big data and: archival theory and practice / construction of memory / provenance**
- **Crowdsourcing** and archives.
- **Specific big data technologies** (e.g. NoSQL databases) and their applications
- **Corpora and reference collections** of big archival data
- **Linked data** and archives
The DCIC is Pursuing a CAS Training / Teaching Agenda

There is a need to:

- create innovative classes that emphasize new modes of collaboration, and interdisciplinary work.
- blend elements of archival thinking and computational thinking:
  - problem solving that uses modeling, decomposition, pattern recognition, abstraction, algorithm design, and scale.
- develop inter-disciplinary iSchools with faculty from Computer Science, Archival Science, and Data Science.
- develop extensive hands-on experience working with cyberinfrastructure to carry out archival functions.

WE WELCOME PARTNERSHIPS -- CONTACT US!
GOAL: Build out the open source DRAS-TIC platform into a horizontally scalable archives framework serving the national library, archives, and scientific data management communities

- **Product** of a 2-year startup by partners, Archival Analytics Solutions Ltd.
- **Scaling** to billions of files and beyond
- **Interfaces:**
  - Web client
  - Command-line client
  - REST storage API (CDMI) industry standard
- **Key-value** metadata
- **Listener** mechanism
- **Python** source on GitHub (Open AGPL license)
- **Apache Cassandra** database (CERN, eBay, GitHub, Hulu, Instagram, Netflix, Twitter…)

WE WELCOME PARTNERSHIPS -- CONTACT US!
Moving forward, build skills in the following areas:

- Digitization
  - Improving image quality
  - Transcription

- Basic Programming
  - Python / Javascript

- Data and Document Management
  - Building collections
  - Digital Storytelling

- Project Management
Questions?

http://DCIC.umd.edu

Email: dcic.ischool@gmail.com