2018-2019 DCIC Project Booklet

Justice, Human Rights, & Cultural Heritage:

1. **Global Journeys, Local Communities** ***-- Ken HEGER***
   A student and professional society-driven project to collect information documenting payment of pensions to American veterans living overseas. The project documents migration patterns, the flow of money, health conditions, and family connections prior to World War I.

2. **Managing White House Correspondents Association (WHCA) Pool Reports**
   -- Richard MARCIANO & Bill UNDERWOOD
   This project is a partnership between the DCIC, UMD's School of Journalism and UMD Libraries. Students will collaborate in creating a digitized <insert text here> Pool Archives.

3. **Legacy of Slavery** ***-- Katrina FENLON***
   In partnership with the Maryland State Archives, two projects using digitized records involving Manumissions and Certificates of Freedom and Underground Railway documentation will mine data and create finding aids to enhance access for users.

4. **Japanese-American WWII Camps** ***-- Richard MARCIANO & Bill UNDERWOOD***
   In partnership with the US National Archives, the project explores the auto-redaction of WWII Japanese-American Camp data. NLP and NER text processing techniques are explored through the development of extraction workflows.

5. **Mapping Inequality** ***-- Richard MARCIANO***
   In partnership with Johns Hopkins, Virginia Tech, and U. of Richmond, a national collection of New Deal *redlining* records is being crowdsourced (these unique records capture racial, ethnic, and economic conditions).

6. **The Human Face of Big Data** **-- Myeong LEE**
   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.

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

7. **Big Data Archives and Analytics** ***-- Greg JANSEN***
   In 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.

8. **Enhancing User Access to Big Data Archives** ***-- Greg JANSEN***
   In partnership with the Medical Heritage Library (a digital curation coalition) students will participate in developing user-friendly interfaces and workflows, and indexes to journals and medical terminology.

★ *Students interested in joining or continuing DCIC projects need to contact Noah Dibert (ndibert@umd.edu) by September 7, 2018.*

1. Global Journeys, Local Communities *-- Ken HEGER*

Goals and Scope
The Global Journeys, Local Communities Project (the Project) will identify, scan and index records documenting the movement of people. It will focus on people who emigrated from the United States to live overseas. These people include American veterans drawing pensions, and other residents of the United States who move abroad and settle in other countries; officials of the U.S. Department of State characterized the latter group as "American colonies."

Sources
In the coming year, the Project will concentrate on records of the U.S. federal government at this stage. Records creators include the Department of State, the Department of the Interior, the Bureau of Pensions, the Department of War, the Adjutant General's Office, St. Elizabeth's Hospital, and Congress. Documents will include correspondence, case files, and reports. The Project will also identify relevant visual images (such as maps, postcards, photographs, and architectural drawings).

Users
The Project will benefit many user groups including
- Micro-biographers (National and International)
- Scholars/Historians (National and International)

iSchool students

How the Project Supports Research
- It will identify and index names of people; micro-biographers will find it of great benefit
- It will produce datasets that people can use to create new knowledge from the documents
- It will scan documents providing remote access to files

How the Project Supports Pedagogy
- The scanned documents provide a repository of documents instructors can use to in workshops, presentations, etc. to illustrate the anatomy of a specific category documents, e.g. a pension file
- It fosters use of computer science techniques and technology to do archival and information management work. For example,
    - Student projects to learn how to scan documents, including how to use different pieces of equipment
    - The scans provide images for students to learn to curate digital images
    - Students will have documents from which they can extract information to create datasets, and link documents through indexes, and various forms of graphs (e.g. trees, graph databases, etc)
- The project is scalable
    - It facilitates the creation of small training modules that iSchool instructors can use in structured classes and in special, short training sessions
    - Small modules are flexible enough to adjust to changing technology and job opportunities

These training modules can be tailored for MLIS, MIM and undergraduate students.

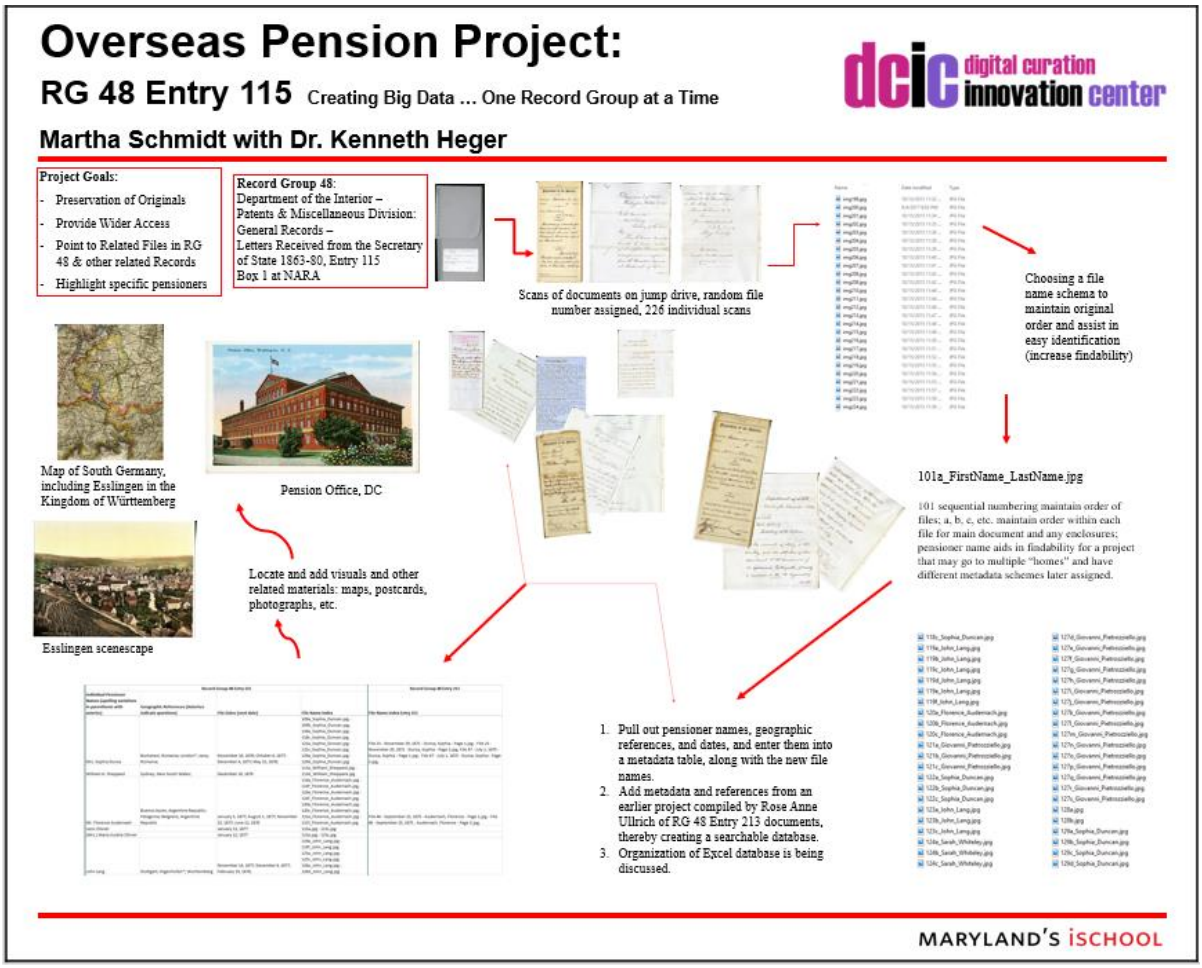
Student Roles and Responsibilities:
Students will work with faculty to process and curate the series of records. They will identify descriptive metadata identifying information such as personal names, geographic locations, military service units, and infirmities (e.g. specific diseases). Students will create descriptive products to foster access to individual records series, to relate documents to other records series, and to populate big data sets.

Deliverables:
Students will work with faculty to identify products appropriate to the work and that foster the student's ability to get a job after graduation. Potential deliverables include posters, blogs, presentations, databases, and articles.

Workload Expectations:
At the beginning of each semester, each student will meet with Dr. Heger, chose a project to expand the data in the Overseas Pension Project, and together they will create a timeline that ensures completion of the project and takes into account the student's work load.

For more information contact – Dr. Kenneth Heger, kheger@umd.edu

The image below was the project poster that was presented at the DCIC Research Showcase at the iSchool in December 2017.



2. Managing White House Correspondents Association (WHCA) Pool Reports

-- Richard MARCIANO & Bill UNDERWOOD

Goals and Scope

This project is a partnership between the DCIC, UMD's School of Journalism and UMD Libraries, and the Newseum Institute. We have an agreement with the White House Correspondents Association (WHCA) to archive and make available digital White House Pool Reports. The main goal of this project is to enhance user access to these documents. www.cbsnews.com/video/saving-presidential-history-new-project-seeks-to-archive-white-house-pool-reports/

What are Pool Reports?

All White House journalists cannot be with the President for all public events, so each day, or for a longer period, a few journalists are scheduled to observe and report the President's activities. These journalists write pool reports and share them with other members of the press pool who are free to edit or use them as they see fit.

Student Roles and Responsibilities:

Students will learn to analyze WHCA press pool reports from the administrations of President George W. Bush and/or President Barak Obama to identify metadata that would be useful in supporting user access to a repository of the pool reports. An existing natural language processing (NLP) tool will be refined to automatically recognize and extract this metadata from the pool reports. A user interface will be designed for using this metadata to support user access to the archived press pool reports.

Deliverables:

Presentation, co-author conference paper, poster

Workload Expectations:

5 – 7 flexible hours per week including regular project meetings. Please reach out to the Project Leads for more information on the schedule for this semester.

Example:

A typical pool report from the Obama Administration, via email, looks like:

> *From: Jordan Fabian <jfabian@thehill.com<mailto:jfabian@thehill.com>>*
> *Date: October 2, 2016 at 4:47:07 PM EDT*
> *To: Tadams-falconer@who.eop.gov<mailto:Tadams-falconer@who.eop.gov>*
> *Subject: In-town pool report #4 - return to WH and lid*
>
> *Motorcade arrived at the White House at 4:43 PM.*
>
> *Pool did not see POTUS exit his vehicle. We also did not catch a glimpse of him on the golf course, per usual.*
>
> *Ride was fairly ordinary as far as presidential motorcades go. Passed by Nationals Park, where the home team was leading the Miami Marlins 3-2 in the bottom of the fourth during the final game of the regular season.*
>
> *We have a travel/photo lid.*

3. Legacy of Slavery

-- Katrina FENLON

Goals and Scope

The Legacy of Slavery in Maryland is a major initiative of the Maryland State Archives. The program seeks to preserve and promote the vast universe of experiences that have shaped the lives of Maryland's African American population. From the day that Mathias de Sousa and Francisco landed in St. Mary's county aboard the Ark and the Dove in 1634, Black Marylanders have made significant contributions to both the state and nation in the political, economic, agricultural, legal, and domestic arenas. Despite what often seemed insurmountable odds, Marylanders of Color have adapted, evolved, and prevailed. The Maryland State Archives, in partnership with the iSchool's Digital Curation Innovation Center (DCIC), have identified two projects appropriate for engagement for iSchool students both graduate and undergraduate.

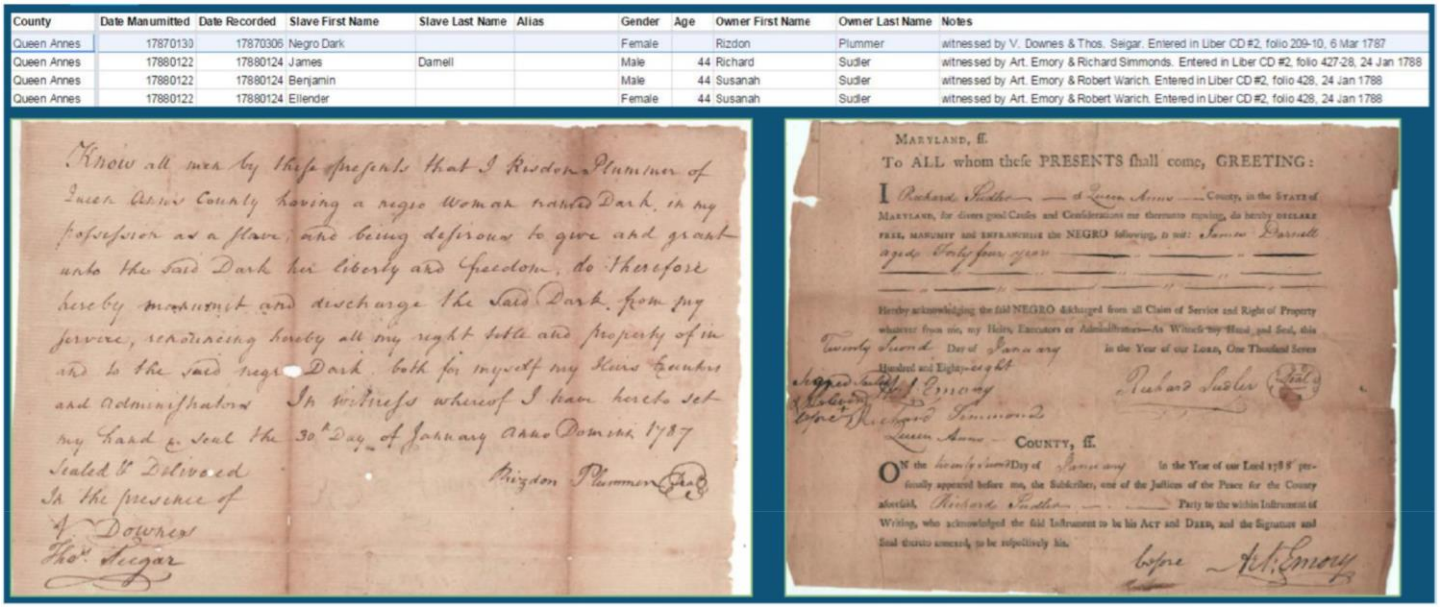
The Maryland State Archives holds two essential types of records documenting freedom within its collection. Manumissions are a legal document that frees an enslaved person from slavery on behalf of the slave holder, and Certificates of Freedom to record proof in county courts of those African Americans born free and those who received freedom from a slave holder. These documents, found in 111 record series created by the counties, contain vital information about those who were enslaved. Included are names and ages of those who were enslaved, names of slave holders, physical descriptions of the enslaved, and the locations of these individuals.

Project 1: Transcription: Certificates of freedom and Census Records in the Maryland State Archives

Students will work with already digitized records (at the Maryland State Archives, the DCIC, or offsite at their own location) to create an item level inventory of Certificates of Freedom records and the Census Records during that time.

Project 2: Data Analytics and Visualization

Students will work with data that is transcribed to make it analytics ready, gather valuable insights from the data and visualize it to enhance access for users. Students will work with state-of-the-art analytics and visualization tools to build interactive dashboards and tell a story from the data.



County	Date Manumitted	Date Recorded	Slave First Name	Slave Last Name	Alias	Gender	Age	Owner First Name	Owner Last Name	Notes
Queen Annes	17870130	17870306	Negro Dark			Female		Rizdon	Plummer	witnessed by V. Downes & Thos. Seigar. Entered in Liber CD #2, folio 209-10, 6 Mar 1787
Queen Annes	17880122	17880124	James	Darnell		Male	44	Richard	Sudler	witnessed by Art. Emory & Richard Simmonds. Entered in Liber CD #2, folio 427-28, 24 Jan 1788
Queen Annes	17880122	17880124	Benjamin			Male	44	Susanah	Sudler	witnessed by Art. Emory & Robert Warich. Entered in Liber CD #2, folio 428, 24 Jan 1788
Queen Annes	17880122	17880124	Ellender			Female	44	Susanah	Sudler	witnessed by Art. Emory & Robert Warich. Entered in Liber CD #2, folio 428, 24 Jan 1788

Work Done to Date

Data present in census records and certificates of freedom have been transcribed into Google spreadsheets. and cleaned to make it analytics ready. Tableau, which is a data visualization tool, has been used to analyze the data and build interactive dashboards to answers specific questions that the Maryland State Archives and other viewers are interested in.

Student Roles, Responsibilities and Takeaways

- Transcribe census records and certificates of freedom into Google spreadsheets.
- Identify interesting questions that should be answered through the transcribed data.
- Explore archival analytical techniques to engage the audience by telling stories from the data.
- Build a data flow pipeline to automate the data cleaning, analytics and visualization processes.
- Participate in project meetings and provide valuable inputs to meet project goal.

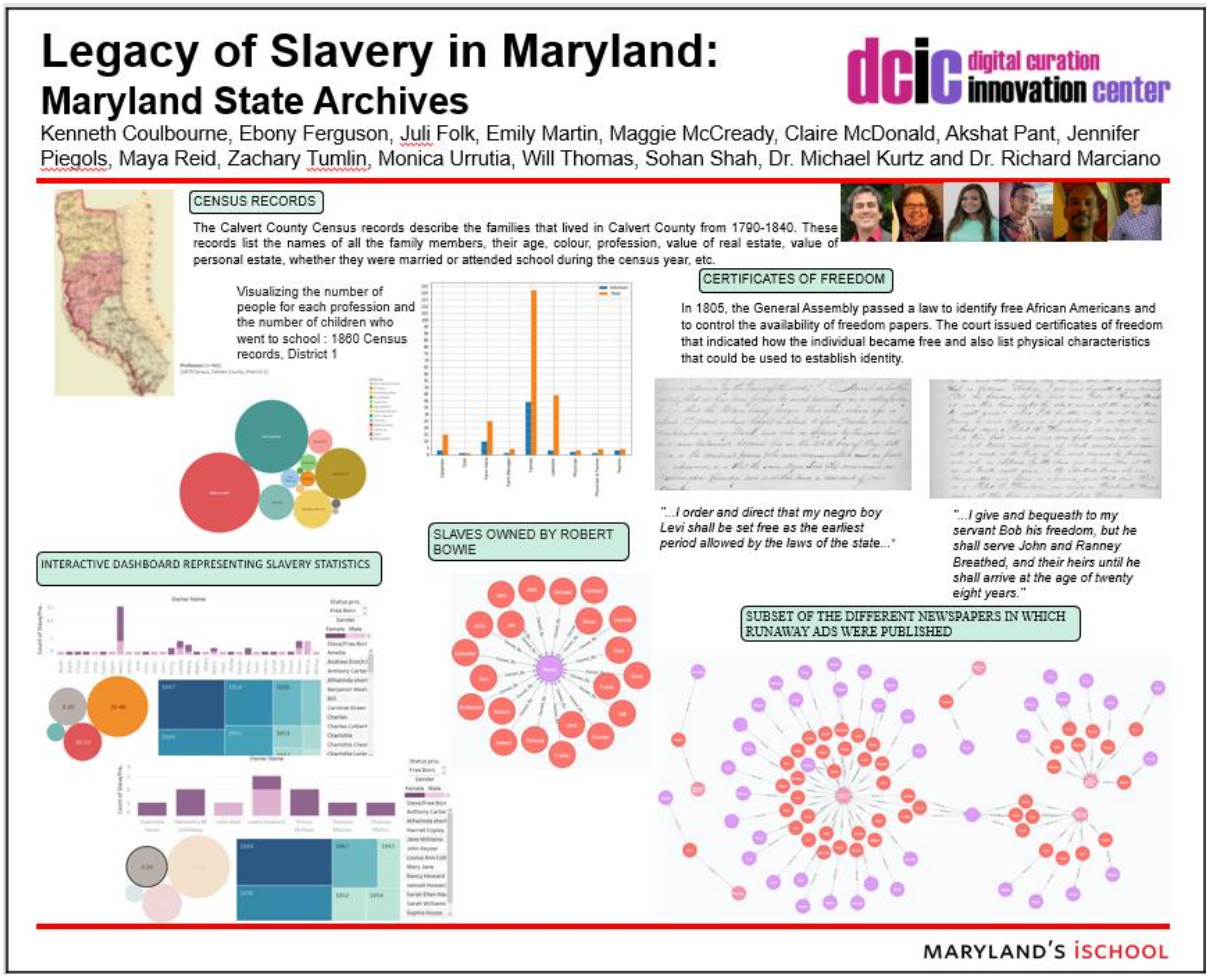
Through this project, students are expected to learn transcribing, data analytics, data visualization, database design and archival analytics.

Deliverables

By the end of the semester, we will transcribe a fixed number of remaining certificates of freedom and census records. We will standardize the transcription process to have accurate data for analysis and automate the data cleaning process. Finally, we will create a plan for publishing our visualizations online for user access.

Workload Expectations

3-5 flexible hours per week.



4. Japanese-American WWII Camps

-- Richard MARCIANO & Bill UNDERWOOD

Goals and Scope

During World War II, over 120,000 Japanese Americans were relocated and jailed across 10 camps. The National Archives maintains record series related to the *War Relocation Authority (WAR)* agency that oversaw the incarceration. Among these are "Internal Security Case Reports" prepared by Relocation Center staff relating to alleged cases of disorderly conduct, assault, theft, loss of property, and accidents. This project focuses on the "Internal Security Cases" index cards. Each card includes a case number, type of charge, names and addresses of persons involved, time and place where the incident occurred, and account of the incident, and refers to a more detailed case file (for which access is restricted). See:

- http://ddr.densho.org/names/
- https://catalog.archives.gov/id/1264228
- http://local.ads.nwsource.com/ads/FlippingBook/2015/Q2/Densho/html/

Work Done to Date

OCR completed for two boxes of incident cards. Attributes like name, birth date and more have been extracted from many of these cards using an NLP software called GATE. A database design has been conceptualized and data from 250 incident cards has been imported into a graph database called Neo4j. Relationships between people, events and households have been identified from this set of cards which has built a platform to import and analyze data from the remaining 25,000 cards.

Students Roles, Responsibilities and Takeaways

- Complete OCR and collect metadata for 2 more boxes.
- Spreadsheet with all item level metadata extracted.
- Data from spreadsheet imported into graph database.
- Identify interesting questions that need to be answered from this data.
- Write cipher queries to visualize relationships in the database.

Through this project, students are expected to learn OCR techniques, use the NLP software GATE to extract data attributes, design a graph database using Neo4j and analyze the data using cipher queries.

Deliverables

By the end of the academic year, we will deliver a demo to NARA's Office of Innovation.

Workload Expectations

5-10 flexible hours per week.

5. Mapping Inequality

-- Richard MARCIANO

Goals and Scope

The 1929 stock market crash devastated America's economy and triggered the beginning of a 10-year economic depression. During this time, American families were at risk of losing homes to foreclosure. To tackle the mortgage crises and restart the Great Depression economy, President Franklin Delano Roosevelt created federal loan programs to refinance troubled residential homes. The US government established the Home Owners' Loan Corporation (HOLC) to determine potential refinance investments by assessing housing and neighborhood conditions. HOLC created maps and area descriptions to describe the features and threats to a particular area; neighborhoods were graded based on the racial/ethnic presence, high and low-income families, and environmental problems. Referring to map shading, grading, and area descriptions, financial institutions made decisions on loan sizes, refinancing opportunities, etc. Unbeknownst to HOLC and the federal government, the 1939 surveys would have major effects on American cities, especially during Urban Renewal in the 1950's. In short, HOLC orchestrated the denial of financial services based on race and ethnic background, better known as redlining. See:

- Newly Released Maps Show How Housing Discrimination Happened (National Geographic, 2016): http://news.nationalgeographic.com/2016/10/housing-discrimination-redlining-maps/
- The Thick Red Line (Terp Magazine, Fall 2016): http://terp.umd.edu/the-thick-red-line/
- T-RACES portal: http://salt.umd.edu/T-RACES/demo/demo.html
- Mapping Inequality (the project's current portal): http://mappinginequality.us

Work Done to Date

Over the past 4 years, teams of archivists have digitized portions of National Archives RG195: General Records of the Home Owners' Loan Corporation [HOLC]. Now that the records are available digitally, students and faculty members are working to curate the collection and "datafy" the information contained within. This involves extracting text from digitized 75-year old records, designing query-able databases to house the information, and disseminating the information so it is publicly accessible. With partners from the University of Richmond, John Hopkins University, and Virginia Tech, the Digital Curation Innovation Center plans to georeference the historical maps and display them on top of online maps.

- Best Poster at the 2017 iSchool Research Showcase
- Best National Geographic Maps of 2016: http://news.nationalgeographic.com/2016/12/best-maps-cartography-2016/

So far, we (1) configured the Scribe platform that allows to crowdsource the transcription of the data, (2) wrote Python scripts to process the Scribe output (JSON) to clean up and organize the data, and (3) designed the relational database that can in-house all different types of HOLC documents.

Student Roles, Responsibilities, and Takeaways

Students will work with colleagues at Johns Hopkins, Virginia Tech, and U. Richmond to develop a big national dataset that will benefit the public and other researchers.

Final Product

National dataset relating to this collection

Skills Exercised

- Digitization
- Optical Character Recognition
- Data Analytics / Data Management
- Geographical Information System referencing
- Python/SQL Programming

Workload Expectations:

Students who are willing to participate in the project are expected to invest about 3 to 5 hours a week on the project consistently. This includes our regular weekly meetings.

6. The Human Face of Big Data -- Myeong LEE

Goals and Scope

The urban renewal project was a nation-wide project to rebuild "blighted" neighborhoods in the 1960s and 70s and resulted in displacing many lively communities. However, the property acquisition processes during the urban renewal period are rarely known not only to researchers, but also to former residents. We aim to build a web-based, big data platform that archives legal documents about property acquisitions during the urban renewal period. Particularly, the platform provides easy-to-use interfaces for navigating property acquisition processes in the parcel level. This application can be used by former residents, archival scientists, and citizens, and ultimately reconstructs digitized neighborhoods where people can see, read, and share their memories.

Work Done to Date

So far, we have achieved (1) implementing the georeferenced map on the web using geographical information systems such as QGIS, ArcGIS, and Leaflet; (2) developing a crowdsourcing platform for curating urban renewal documents; (3) designing databases and search interfaces through data modeling and iterative design approaches, (4) publishing two conference papers in IEEE Big Data 2017 and iConference 2018, and (5) conducting interviews with potential stakeholders to better design the system. For a project description, see: https://www.youtube.com/watch?v=DUKcNcJvOik

Student Roles, Responsibilities, and Takeaways

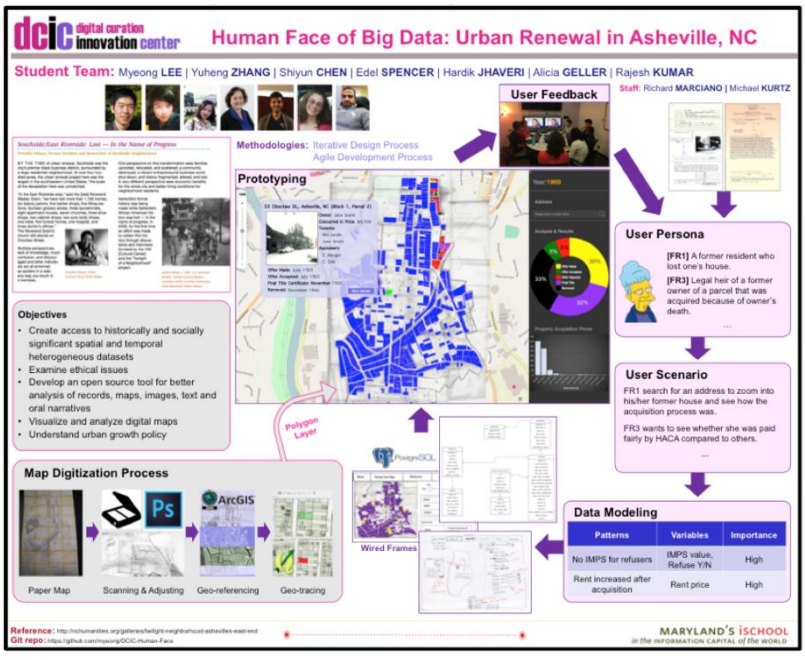
We are in the stage of (1) digitizing urban renewal documents and finding data patterns, (2) refining the web interfaces for providing user-centered interfaces; and (3) taking a value-sensitive design approach for designing a historically-sensitive data platform. Students who are participating in this project are responsible for digitizing documents, assuring data quality, conducting interviews with potential users, and analyzing interview data. Skillsets and qualifications that students are expected to learn from the project are digital curation, web development, data management, and qualitative research.

Deliverables:

We expect to interview potential users for better designs, and to launch the urban renewal archive as a web application with a complete dataset in the database.

Workload Expectations and Application:

In Fall 2018, we will only recruit 5 new members. Due to the spot limitation, students need to send their resumes to Myeong Lee (myeong@umd.edu) with a brief description explaining why they want to participate in the project. Students willing to join this team are expected to be proactive and work for about 5 to 7 hours a week.



7. Big Data Archives and Analytics *-- Greg JANSEN*

This initiative encapsulates two projects – Brown Dog and DRAS-TIC Fedora.

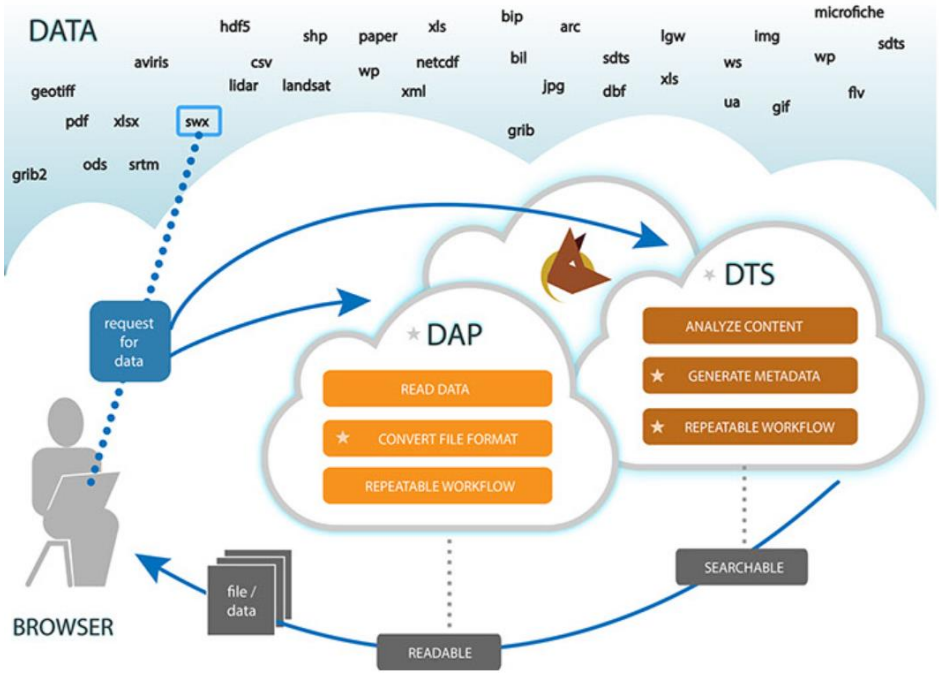
7.1. Brown Dog

Goals and Scope:
Brown Dog seeks to develop a service that will make past and present un-curated data accessible and useful to scientists while also demonstrating the novel science and scholarship that can be conducted from such data.

Brown Dog will not attempt to construct a single piece of software that magically understands all data, but instead will use every possible source of automatable help already in existence (e.g. software, tools, libraries, other services) in a robust and provenance preserving manner to create a service that can deal with as much of this data as possible. Brown Dog is the proverbial "super mutt" of software, serving as a low-level data infrastructure to interface with digital data content across the web and enabling a new era of science and applications at large. The broader impact of this work is in its potential to serve not just the scientific community but the general public as a "DNS for data", transforming data on the fly to more accessible forms through a distributed and extensible collection of data manipulation tools, moving civilization towards an era where a user's access to data is not limited by a file's format or un-curated collections.

Student Roles and Responsibilities:
Contribute useful new software tools to Brown Dog's auto-curation cloud services. (http://browndog.ncsa.illinois.edu/) You will be introduced to key technologies, such as Docker containers, Python, JSON-LD, and high performance messaging systems. Each individual or small team will build a useful file converter or metadata extractor that will become your personal contribution to the Brown Dog cloud service. We have identified several such tool needs that are particular to the archival use of Brown Dog, so your work will directly benefit the archival community. Students will need to have some basic software development experience with Python or Java, but will be introduced to the rest.

Workload Expectations:
5 - 10 flexible hours per week



7.2. DRAS-TIC Fedora

Goals and Scope:
IMLS funded 2-year project for research and development of Fedora repository-compatible software that achieves web scale and is open to computational approaches to improving collections and access. Various technical and non-technical student roles are available in this large software development and testing project.

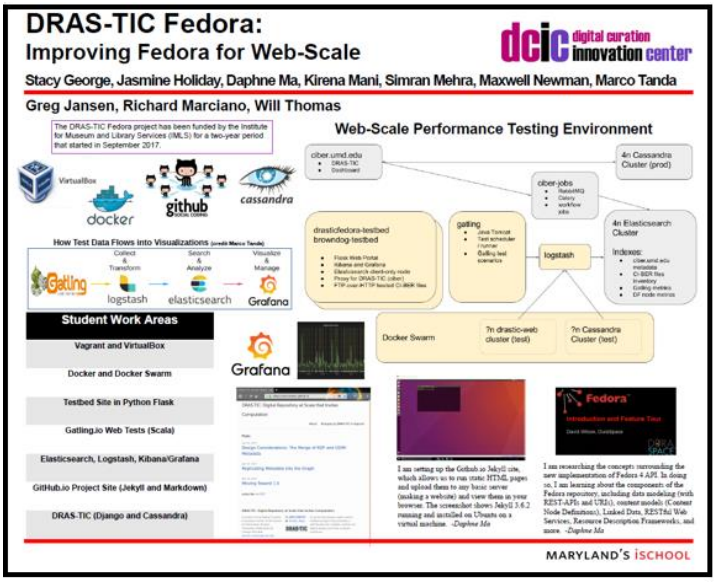
Work Done to Date:
The Fedora-compatible digital repository software will be based on existing software, DRAS-TIC, which provides many of the services required by the Fedora specification. A new core persistence service, based on the Link Data Platform (LDP) specification is prototyped and ready for testing. When it is proven scalable, we will modify the DRAS-TIC software to use the new persistence service to store data. We also completed a use case analysis for our four partner institutions. These use cases are to be the basis for our testing regime. The core data service will also need extensive testing for LDP API compliance.

Student Roles and responsibilities:
Much of the project time will be spent on creating informative automated software tests. Analysis of test results and gap analysis versus use cases will also be a major focus. Students can become involved in a technical capacity in the software project. We welcome help with software testing, and even software development. We also welcome student participation through the broader computational archives project. Please see that flyer for details.

Skills Offered:
- reading technical specifications
- writing scenario-based software requirements
- organizing and balancing stakeholder needs
- agile software methodology
- software quality assurance

Deliverables:
Students will need to contribute code, data, and technical guides to Git code repositories as appropriate for their work. They will evaluate software test results as needed and create short reports on outcomes of testing.

Status
The project is primarily a DCIC staff initiative and we are not accepting students at this time.



8. Enhancing User Access to Big Data Archives *-- Greg JANSEN*

Goals and Scope:

The Medical Heritage Library (MHL) is a digital curation collaborative among some of the world's leading medical libraries to promote free and open access to quality historical resources in medicine and the human health sciences. The goal is to provide the means by which readers and scholars across a multitude of disciplines can examine the interrelated nature of medicine and society, both to inform contemporary medicine and strengthen the understanding of the world in which we live. Students will leverage ArchiveSpark to make access to these collections more intuitive to users.

What is ArchiveSpark?

ArchiveSpark is a Java/JVM library, written in Scala, which can be used as an API in any Java/Scala/JVM program for easy and efficient access to Web archives and other supported datasets. In addition to that, it can be used stand-alone using Scala's interactive shell or notebook tools, such as Jupyter.

Student Roles and Responsibilities:

1. Make ArchiveSpark with MHL more intuitive by developing a user-friendly interface (or other mechanism) for making ArchiveSpark functionality more broadly accessible. This project seeks to make ArchiveSpark workflows broadly accessible to the public. Products of this project could include creating a number of canned recipes for searching content with ArchiveSpark and considering new approaches to searching the dataset for the purpose of extraction and analysis easier for researchers.

2. Connect Index cat to journal articles that have been digitized by the MHL. This challenge involves matching Index Cat entries with full text articles residing in the Medical Heritage Library.

3. Create an index of archaic medical terminology using medical dictionaries found in the Medical Heritage Library, map those terms to contemporary medical terminology (such as the Unified Medical Language System, and index the Medical Heritage Library corpus to facilitate the discovery of published content from the perspective of contemporary medicine.

Workload Expectations:

5 - 10 flexible hours per week