IJDC | *General Article*

Building Open-Source Digital Curation Services and Repositories at Scale

Richard Marciano
University of Maryland iSchool

Gregory Jansen
University of Maryland iSchool

Will Thomas
University of Maryland iSchool

Sohan Shah
University of Maryland iSchool

Michael Kurtz
University of Maryland iSchool

Abstract

The focus of this article is to share several in-progress research and development open-source approaches that seek to design, build, and test digital curation services and repositories that have the potential to scale (the IMLS-funded Fedora DRAS-TIC and the NSF-funded Brown Dog). We also discuss the creation of a big records testbed of justice, human rights, and cultural heritage collections (100 TB and 100 million records), the emergence of Computational Archival Science (CAS), and the resulting efforts at integrating digital curation education and research.

We ultimately seek to develop a sustainable community of users and developers, with solutions that serve the international library, archives, and scientific data management communities. We are also focused on digital curation training and education in these innovative environments.

Received 22 January 2018 ~ *Accepted* 20 February 2018

Correspondence should be addressed to Richard Marciano, 4130 Campus Drive, 4110C Hornbake Library South, College Park, Maryland 20742-4345. Email: marciano@umd.edu

An earlier version of this paper was presented at the 13th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

http://dx.doi.org/10.2218/ijdc.v13i1.621
DOI: 10.2218/ijdc.v13i1.621

Introduction

We present two approaches to the design and curation of digital repositories that exploit current technology to address the emerging issues of capacity scaling, heterogeneous content, and sustainability:

- The first approach exploits NoSQL distributed database technology to support repositories that can scale out horizontally to thousands of commodity servers. This was recently funded through a U.S. Institute of Museum and Library Services (IMLS) grant, called DRAS-TIC Fedora[1], as part of IMLS's National Digital Platform (NDP) program.

- The second approach exploits web-scale server virtualization to support a curation service, known as Brown Dog[2]. This is an ongoing U.S. National Science Foundation (NSF) DiBBS-funded project (Data Infrastructure Building Blocks). This service provides web and API access to hundreds of tools, as created by our partners at the National Center for Supercomputing Applications (NCSA), and is employed in our own archival repository case study.

These two approaches combine to add value to large digital collections. Brown Dog enables extractor and converter filters to be applied to workflows for appraisal and ingestion. These filters generate metadata and create surrogates for preserved records. In addition to OCR and detection of application type (e.g. recognizing that a record is a PDF file), natural language processing (NLP) extractors performing named-entity recognition (NER) can be added to the workflow to create more useful description metadata (Jansen, Marciano, Padhy and McHenry, 2016).

DRAS-TIC Fedora's ability to scale to large numbers of objects means that we can use collections as corpora for training and testing machine learning (ML) approaches to NLP, such as relationship extraction; trained recognizers can be added as Data Tilling Service (DTS) filters in Brown Dog. Combined, these two approaches will not only scale in the number of records they can describe and convert, but will be able to characterize and describe those records in greater individual detail and, importantly, improve on the accuracy of those filters using their combined resources. We think that these two technologies, distributed databases and web-scale virtualization, are key factors that will make repositories and their workflows sustainable in the future as they continue to grow toward the petabyte scale with even greater diversity of content.

Finally, we discuss the creation of a testbed of justice, human rights, and cultural heritage collections, the emergence of Computational Archival Science (CAS), and the resulting efforts at integrating digital curation education and research.

[1] DRAS-TIC Fedora: http://dcicblog.umd.edu/dras-tic-fedora/

[2] Brown Dog: http://browndog.ncsa.illinois.edu/

Development of Distributed Scalable NoSQL Catalogs and Repositories (DRAS-TIC)

The Digital Curation Innovation Center (DCIC)[3] at the University of Maryland's iSchool is currently researching, developing, and testing software architectures to improve the performance and scalability of the Fedora repository.[4] This project explores the creation of a new Fedora implementation without current performance bottlenecks relating to storage size, enabling institutions to manage Fedora repositories with petabyte-scale collections. It applies the new Fedora 5 application programming interface (API) to the DCIC's open-source repository software stack called DRAS-TIC[5]. This will bring the benefits of DRAS-TIC to the Fedora community, fully supporting all Fedora 5 compatible software, including existing websites and workflow tools. DRAS-TIC (Jansen and Marciano, 2016) provides scalable, fault-tolerant object storage, built on the Cassandra NoSQL distributed database. Partners include Fedora (Leadership Group and Steering Committee), Smithsonian Institution (Office of Research Info. Services and National Museum of American History), University of Illinois Urbana-Champaign National Center for Supercomputing Applications (NCSA), University of Maryland Libraries, and Georgetown University Library. These partners are helping us develop use cases and performance expectations. The project is expected to produce open source software, tested system configurations, documentation, and best-practice guides. DRAS-TIC is an acronym that stands for digital repository at scale that invites computation (to improve collections).

The promise of DRAS-TIC is that memory institutions can incrementally grow a fully-functional repository as their collections grow, instead of having to forklift in new enterprise storage, perform massive data migrations, and face performance bottlenecks and single points of failure that stem from vertical (centralized) storage strategies. We find that big, centralized repositories create longer and more expensive planning cycles that dramatically inhibit new collection development. In contrast, adding capacity to a distributed repository comes with predictable effort and marginal costs. These benefits, known as horizontal scaling, have driven web-scale businesses to rely more and more upon distributed storage. Apache Cassandra in particular, originally developed by Facebook, has been adopted by a long list of companies, including Apple, Netflix, eBay, and Microsoft. DRAS-TIC also has a workflow module, based on message queues that are also persisted in Cassandra, so that workflow tasks can smoothly scale up alongside the object storage. This project brings key best practices that have matured in industry into the Fedora community.

[3] Digital Innovation Curation Center (DCIC): http://dcic.umd.edu

[4] Improving Fedora to Work with Web-scale Storage and Services: https://www.imls.gov/sites/default/files/grants/lg-71-17-0159-17/proposals/lg-71-17-0159-17-full-proposal-documents.pdf

[5] Digital Repository At Scale – That Invites Computation (That Improves Collections) – DRAS-TIC: https://github.com/UMD-DRASTIC/drastic

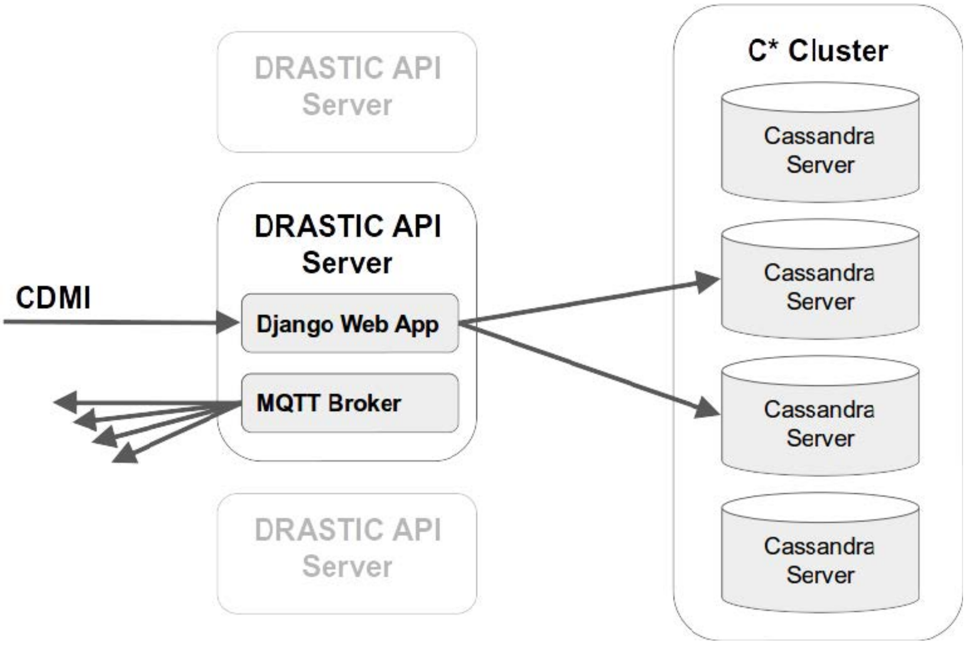

Figure 1. DRAS-TIC's front-end and back-end can scale out.

Usually we think of workflow in reactive terms, as a series of tasks performed on objects upon ingest or some other trigger, taking each object individually and adding value somehow. However, in addition to this approach, supported by the DRAS-TIC workflow engine, there is the notion of analytics and corpus or collection-level processing. Analytics instead takes an entire collection or an entire repository as a dataset. This idea has been gaining currency in the past year as the 'Always Already Computation' initiative has promoted the understanding of Collections as Data[6]. In the DCIC we have also witnessed an explosion in corpus-level analysis of collections, from natural language processing (NLP) to image recognition and more. Researchers in computational archival science now reach for machine learning as a practical tool, training their models on as much data as possible to improve accuracy. Computational approaches to collections are more efficient when they can read a lot of objects quickly, often simultaneously, an access pattern that is ideally supported by a distributed repository like DRAS-TIC.

In addition, it is maximally efficient to run certain modes of analysis natively, across the Cassandra cluster itself, with the data in place. The DRAS-TIC data schema in Cassandra was designed to facilitate this form of massively parallel analysis, through computational tools like Apache Spark. Spark is a parallel compute framework that allows one to stream DRAS-TIC objects and metadata as normal datasets into our own unique functions or into off the shelf algorithms, such as the Spark machine learning library. As the collections as data approach matures, we hope to incorporate the more routine analytical functions into a repository analytics module, providing common functions as well as examples for those developing their own algorithms.

The DRAS-TIC Fedora project, funded by a two-year National Digital Platform grant from the IMLS, is producing open-source software, tested cluster configurations, documentation, and best-practice guides that enable institutions to reliably manage Fedora repositories with petabyte-scale collections.

[6] Collections as Data: http://digitalpreservation.gov/meetings/asdata/impact.html

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

Brown Dog is a $10.5M NSF/DIBBs-funded collaboration with the University 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. Brown Dog is a web service, hosted at the NCSA, that addresses the proliferation of heterogeneous formats entering digital archives. The Brown Dog service is composed from a long list of third party software tools that have been packaged in virtual machines accessible by a common REST API. The service applies any tools it can find to either extract metadata from your input file or convert your file into your designated format. In this way Brown Dog allows us to address the 'long tail' of file formats, including those minor formats that rarely rise to prominence within any one repository.

The Brown Dog service can help to transform legacy files that are functionally opaque to modern desktop software, into extracted metadata and recent or more standard file formats that are still well-supported and thereby improve ongoing access to the intellectual content.

Despite a broad collection of tools, Brown Dog does not yet serve every format imaginable. Therefore the service is designed to facilitate third-party contribution of new tools. Institutions using Brown Dog can implement a new local workflow tool, or they can package the tool for use by all institutions that use Brown Dog. These contributed tools free other organizations to focus their capacities on the content in their collections rather than having to implement a tool to access or describe those formats.

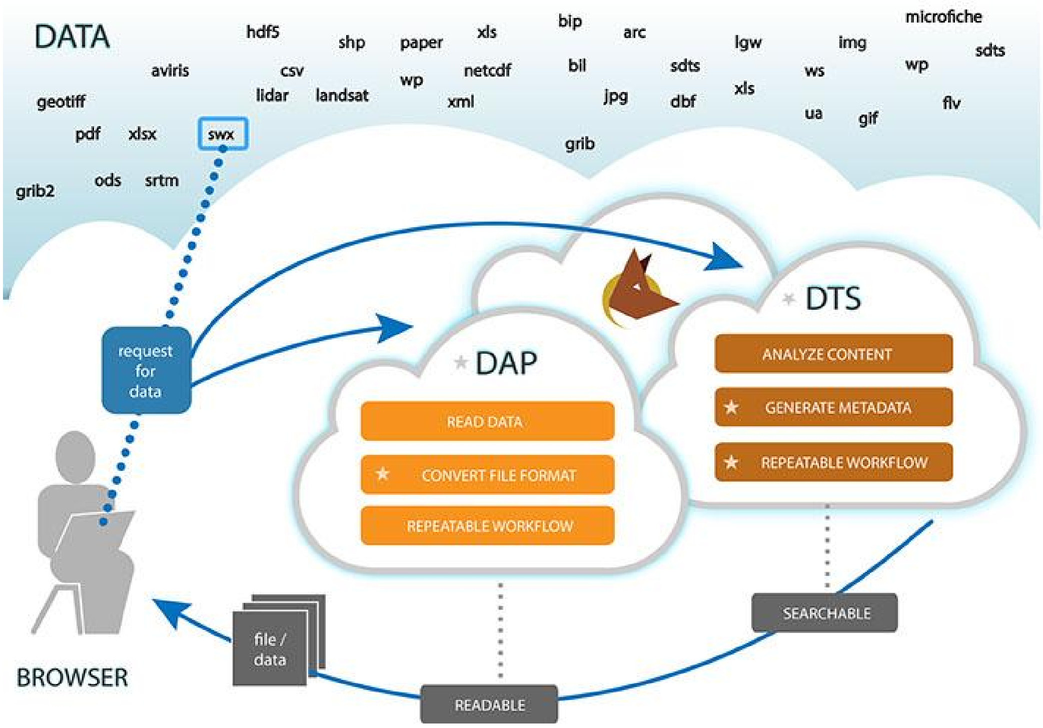

Figure 2. Brown Dog's service framework.[7]

[7] Brown Dog – How it Works: http://browndog.ncsa.illinois.edu/#how_it_works

The DCIC provides archival use cases that help steer Brown Dog's service development. Our software architect is embedded with the Brown Dog development team on sprints and takes on development tasks that have an archival focus. Along with partners in the archives community, we created a wish list of additional conversion and extraction tools, including some experimental approaches, that we find useful for archival workflows. Our students and staff work together on small projects to prototype and then contribute tools to the Brown Dog service. Lastly, we are deeply engaged in web-scale performance testing of the service, feeding it file data sampled from the 100 million files in the data testbed. Our tests include stress tests, to ensure that performance does not degrade under web-scale load, and qualitative tests of the services' response to diverse file formats.

We encourage the reader to consider the efficiencies that the Brown Dog service creates for processing heterogeneous collections. Those looking for more information may wish to visit the project website and try the service with your own sample data.

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

We are currently testing these approaches with cultural heritage archival collections. These include a number of unique collections from 18th and 19th Century US Slavery records, US New Deal redlining records on racially zoned city neighborhoods, World War II Japanese-American Camp records, and 1960-1970s urban renewal housing records.

Projects include both justice, human rights, and cultural heritage themes (community displacement, racial zoning, refugee narrative, citizen narrative, movement of people, and revealing untold stories) and cyberinfrastructure for the curation and management of digital assets at scale themes (preservation services in the cloud, and scalable distributed repositories).

These projects[8] are supported by the development of the DRAS-TIC open-source software which currently manages 100 million files and 100TB of cultural heritage data. This testbed grew out of previous NSF/US National Archives (NARA) supported research and was assembled with the support of NARA's Applied Research staff. Other than the collections mentioned above, the bulk of the 100TB are electronic records from over 150 federal agencies. They exemplify highly heterogeneous content with over 6,000 file types ranging from a few files to tens of millions of files each, and including diverse file types (text, desktop publishing, databases, audio, video, GIS, XML, etc.) and historical, cultural social science, and scientific content. The testbed had primarily been used to support the development of scalable record visualization of e-records and to test the development of national federated infrastructure.

The 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] Practical Digital Curation Skills for Archivists in the 21st Century: https://drum.lib.umd.edu/handle/1903/18865

Exploration of a New Trans-Disciplinary Field: Computational Archival Science (CAS)

We finally make a case for integrating all these educational and research activities into a digital curation center environment (Digital Curation Innovation Center – DCIC), where advances in computational treatments of archival and cultural content are promoted, through a new trans-disciplinary field we call CAS or Computational Archival Science. We believe the emergence of CAS to follow advances in Computational Social Science, Computational Biology, and more recently Computational Journalism, defined as the "finding and telling news stories, with, by, or about algorithms" (Diakopoulos, 2016).

We define CAS as an emerging activity 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. This suggests that computational archival science is a blend of computational and archival thinking. See our CAS portal for the latest update on these developments.[9]

A good example of CAS can be found in the Smithsonian Institution's use of machine learning to categorize five million museum botanical specimens.[10]

Over the last year a number of initiatives been launched with computational treatments of library collections. The 'Always Already Computational' IMLS-funded project[11] is exploring the impact of computationally-driven research and teaching. The Library of Congress has started a new group call 'National Digital Initiatives (NDI)' where Collections as Data[12] is the focus. In the UK, an AHRC-sponsored workshop in June 2017 at the British Library has identified challenges and opportunities for managing big data in the heritage sector (Harrison, Morel, Maricevic, and Penrose, 2017).

Finally, at the University of Maryland's DCIC Center, an initiative called CAS is exploring computational treatments of archival and cultural content. The founding partners include researchers from the University of Maryland (Richard Marciano, Bill Underwood, Michael Kurtz, and Greg Jansen), Canada (Victoria Lemieux at UBC), the UK (Mark Hedges at King's College London), the University of of Texas (Maria Esteva from TACC), and the US National Archives (Mark Conrad). A foundational book chapter on CAS called 'Archival Records and Training in the Age of Big Data', is to be published in May 2018 (Marciano et al., 2018). It explores eight topics: 1) Evolutionary prototyping and computational linguistics (Bill Underwood), 2) Graph analytics, digital humanities 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), and 8) Spatial and temporal analytics (Mark Conrad).

[9] Computational Archival Science (CAS) Portal: http://dcicblog.umd.edu/cas

[10] 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/

[11] Always Already Computational: https://collectionsasdata.github.io

[12] Collections as Data: http://digitalpreservation.gov/meetings/asdata/impact.html

Table 1. Papers presented at the CAS#2 Workshop at the IEEE Big Data 2017 conference in Boston on December 13, 2017, showing how archival concepts are matched with computational methods.

Project Name	Archival Concepts	Computational Methods
A. Exploring Archival Data		
#1: Building new knowledge from distributed scientific corpus [France & Netherlands]	Trusted digital repositories (TDR), digitization, cultural heritage platforms	EUDAT automated scalable e- infrastructure, integrated computational services, document scanning, OCR
#2: An Infrastructure and Application of Computational Archival Science to Enrich and Integrate Big Digital Archival Data [Taiwan]	Big archival data	Record linking, GIS
#3: Computational Curation of a Digitized Record Series of WWII Japanese-American Internment [USA]	Digital curation, automated metadata extraction	NLP, NER, GIS, Graph database, linked data
#4: The Cybernetics Thought Collective Project [USA]	Geographically dispersed archives	NLP, NER, machine learning
B. Curation and Appraisal		
#5: Towards Automated Quality Curation of Video Collections from a Realistic Perspective [USA]	Collection assessment, quality-aware metadata for video collections to inform appraisal, preservation, and access decisions, quality detection in videos	Feature computing from video records, automated quality prediction, scalable HPC
#6: Line Detection in Binary Document Scans [USA]	Digitization, Classification of archival images	Line detection, image segmentation, OCR
#7: Auto-Categorization & Future Access to Digital Archives [Canada & USA]	Recordkeeping, Record disposition	Auto-categorization, document- classification, machine learning
#8: Heuristics for Assessing Computational Archival Science (CAS) Research [USA]	Records of an urban renewal project (property documents, map)	Heuristics for CAS research, Digital Curation (scanning and adjusting, geo-referencing, geo-tracking), System Design (interface design, database design), Implementation (prototype), Value-sensitive design

Continued over/-

Project Name	Archival Concepts	Computational Methods
C. CAS Methodologies		
#9: What Can a Knowledge Complexity Approach Reveal About Big Data and Archival Practice? [Netherlands]	Knowledge complexity in archives, requirements for working with complex collections, context, resistance to computational approaches	Big data analytics, data science
#10: Protecting Privacy in the Archives [Canada]	Personally Identifiable Information	NLP, NER, sentiment analysis, topic modeling
#11: Identifying Epochs in Text Archives [United Kingdom]	Classification of time- coded collections of textual collections into epochs and periods	Cultural analytics, topic modeling
#12: GraphQL for Archival Metadata [United Kingdom]	Query interfaces to archival materials	APIs for cultural heritage materials, graph databases, structured data query
D. Creation and Management of Current Records		
#13: The Blockchain Litmus Test [USA]	Decentralized record keeping	Blockchain, secure computing, trustworthiness, risk analysis
#14: A Typology of Blockchain Record keeping Solutions and Some Reflections on their Implications for the Future of Archival Preservation [Canada]	Record keeping, digital preservation, archival trust	Blockchain, computational validation, distributed ledger, computational trust

Table 1 demonstrates the emerging CAS blending elements of archival thinking and computational thinking, a form of problem solving that uses modeling, decomposition, pattern recognition, abstraction, algorithm design, and scale (Wing, 2006).

Integration of Digital Curation Education and Research

The overall objective of the DCIC is to promote digital curation education and training through innovative instructional design (offered in-person, and online in fall 2018), integrated with student-based project experience. This is the building block for research in building open-source digital curation services and repositories at scale. The key components of this initiative include:

- Creating a new academic specialization (Archives and Digital Curation) in the Masters of Library and Information Science program, to prepare students for careers in the information field of the 21st century;
- Developing a series of courses for graduate and undergraduate students in the iSchool that teach digital curation theory and practice through lectures, discussions, readings, and in-depth experience on team-based, hands-on digital curation projects led by senior DCIC faculty and staff;

- Organizing seminars for graduate students to define the theoretical and operational elements of Computational Archival Science. This provides participants the opportunity to explore how computational and archival thinking can be applied to the complex issues confronting the management and preservation of repositories at scale and developing digital curation services;
- Establishing a Digital Curation for Information Professionals (DCIP) Certificate program. This is a three-course, fully online program designed for working professionals who need training in next generation cloud computing technologies, tools, resources, and best practices to help with the evaluation, selection, and implementation of digital curation solutions;
- Offering students participation on interdisciplinary digital curation projects with the goal of developing new digital skills, conducting interdisciplinary research at the intersection of archives, digital curation, Big Data, and analytics. One example is with the MD State Archives' Legacy of Slavery project[13].
    - iSchool students work with digitized census records and Certificates of Freedom to crowdsource data (personal information about African Americans born free and those freed by a slaveholder), and then apply graph database and visualization technologies to create links and relationships leading to the telling of untold stories.

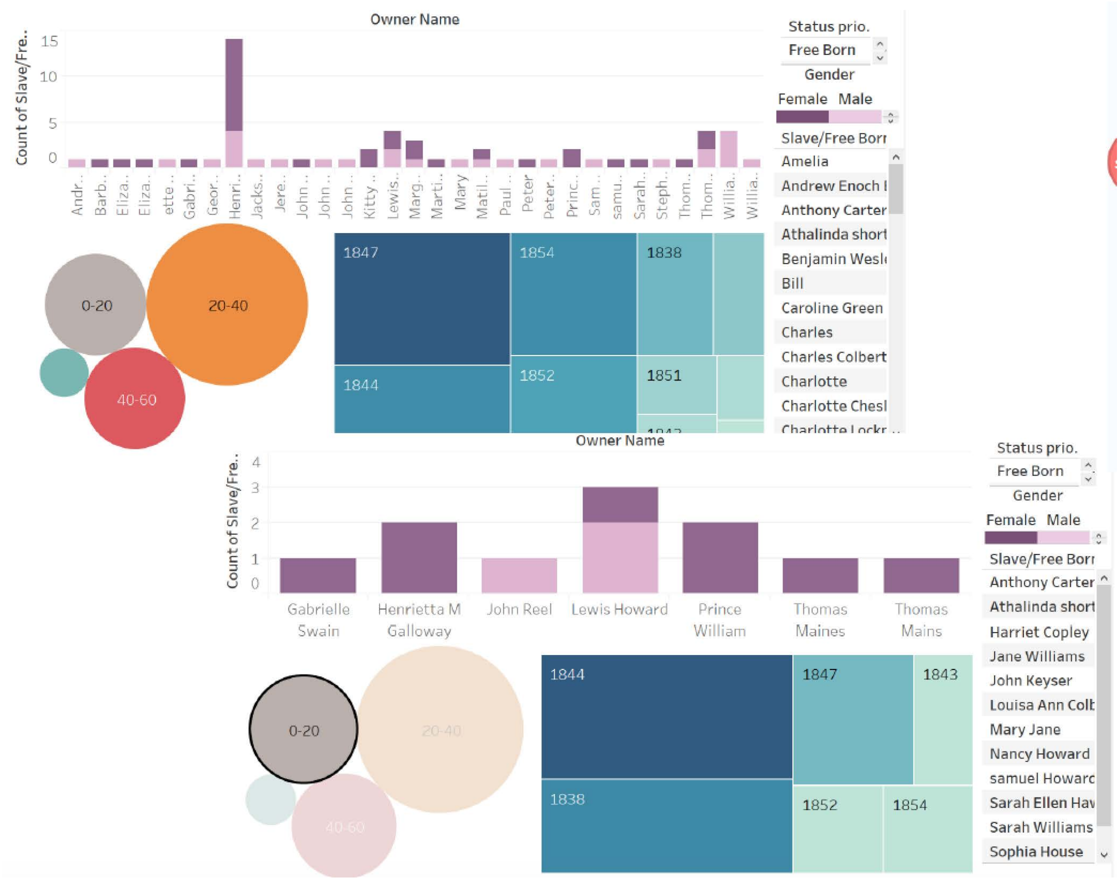

Figure 3. Interactive Tableau dashboard representing slavery statistics.

[13] Legacy of Slavery project: http://dcicblog.umd.edu/legacyofslaveryinmaryland/

Additional visual analytics using graph databases can be found in the Appendix.

Acknowledgements

We wish to acknowledge ongoing support from Maryland's iSchool to the Digital Curation Innovation Center (DCIC) and also major funding from the National Science Foundation (NSF) and the Institute of Museum and Library Services (IMLS). The Brown Dog project (NSF Cooperative Agreement ACI-1261582), is a $10.5M NSF/DIBBs-funded collaboration with the University of Illinois at Urbana-Champaign NCSA Supercomputing Center. The work in this paper has leveraged the Brown Dog funded dataCave and DRAS-TIC open-source platform (Digital Repository At Scale – That Invites Computation). The DRAS-TIC Fedora project (IMLS LG-71-17-0159-17) aims at improving the Fedora repository to work with web-scale storage and services.

References

Diakopoulos, N. (2016). Algorithmic transparency in digital curation. Paper presented at the CAS Symposium, 2016. Retrieved from https://drive.google.com/file/d/0B9kwFSGeIVm8RGJfWVBoYWpnTjQ/view

Harrison, R., Morel, H., Maricevic, M., & Penrose, S. (2017). Heritage and Data: Challenges and Opportunities for the Heritage Sector. Report of the Heritage Data Research Workshop at the British Library. Retrieved from https://heritage-research.org/app/uploads/2017/11/Heritage-Data-Challenges-Opportunities-Report.pdf

Jansen, G., Marciano, R., Padhy, S., & McHenry, K. (2016). Designing scalable cyberinfrastructure for metadata extraction in billion-record archives. Paper presented at iPRES 2016, Basel, Switzerland.

Jansen, G. & Marciano, R. (2016). DRAS-TIC measures: Digital repository at scale that invites computation (to improve collections). Paper presented at CNI Fall 2016 in Washington D.C. Retrieved from https://www.cni.org/topics/digital-curation/drastic-measures-digital-repository-at-scale-that-invites-computation-to-improve-collections

Marciano, R., Lemieux, V., Hedges, M., Esteva, M., Underwood, W., Kurtz, M., & Conrad, M. (2018). Advances in librarianship – Re-envisioning the MLIS: Perspectives on the future of library and information science education. Retrieved from http://dcicblog.umd.edu/cas/wp-content/uploads/sites/13/2016/05/Marciano_Kurtz_et-al-Archival-Records-and-Training-in-the-Age-of-Big-Data-final-1.pdf

Wing, J. (2006) Computational thinking. *Communications of the ACM, 49*(3), 33-35. Retrieved from https://www.cs.cmu.edu/~15110-s13/Wing06-ct.pdf

Appendix

Graph Database (Neo4j) Visualizations of Slave and Owner Relationships

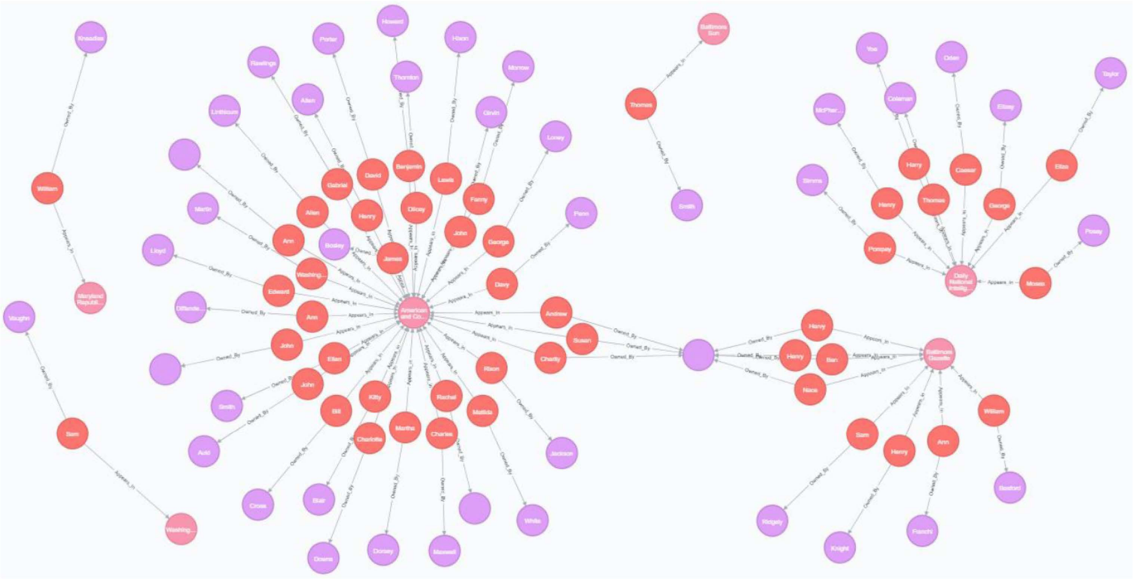

Figure 4. This figure shows a subset of the newspapers in which Runaway Slave ads were published. Pink nodes are newspapers, red nodes are 'Slave' names, and purple nodes are 'Owner' names. The relationships are 'Appears_In', and 'Owned_By'.

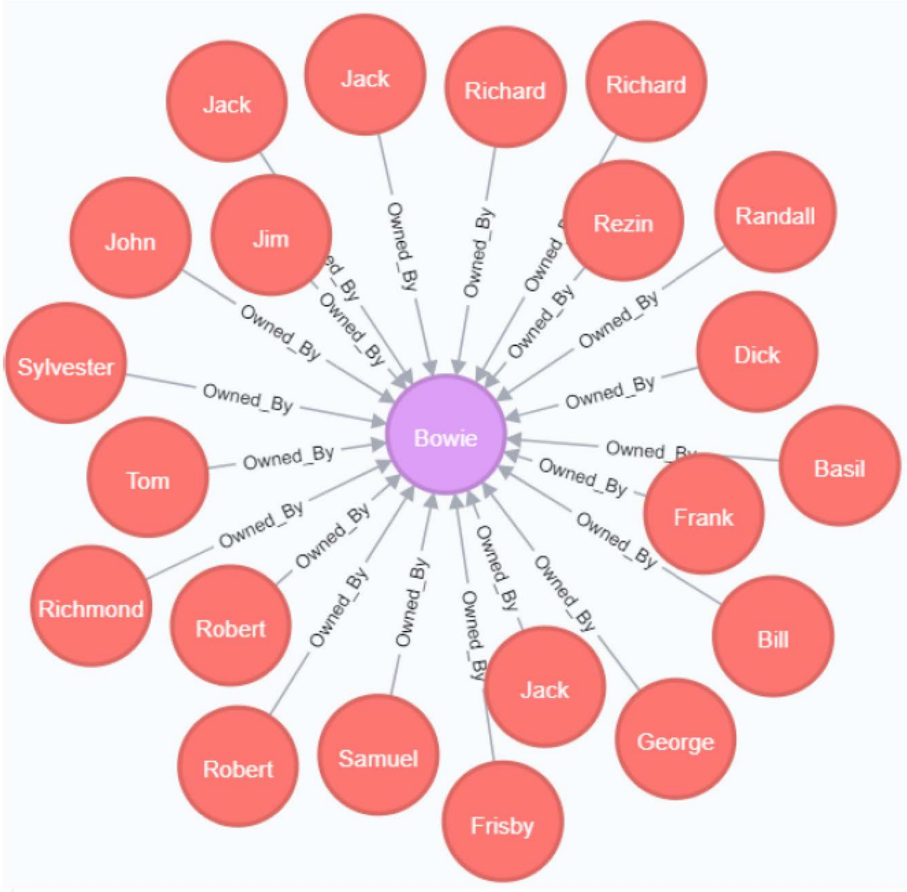

Figure 5. This figure shows which slave names pertain to 'Robert Bowie'. The purple node is the 'Owner', the red nodes are 'Slave' names. The relationship is 'Owned_By'.