Scaling Up: Machine-Assisted Description of Historical Materials

Proposal for Amazon Research Award
Principal Investigator (PI): Christopher J. Prom
Co-PIs: Bethany Anderson, Patricia Lampron, Tom Habing, and Kyle Rimkus.
September 15, 2017

Letters, memoranda, diaries, reports, scrapbooks, photographs, audio recordings, and films: Such one-of-a-kind documentary materials record the history and preserve the memory of people, families, communities, organizations, states, nations, and the global community. With easy access to these sources, people from all walks of life can understand and learn from the past, as well as work toward a better tomorrow. The project described in this proposal—*Scaling Up: Machine-Assisted Description of Historical Materials*—aims to significantly improve and transform access to these materials by 1) applying natural language and machine learning tools to a large body of such records from the Library at the University of Illinois at Urbana-Champaign and 2) testing an AWS-based service for potential implementation with other universities, archives, and libraries.

Problem Statement, Literature Review, and Rationale

Archival records are typically not easy to find through services like Google, Amazon, or Facebook, three platforms that many people rely on for their information needs. Instead, people locate them in national, state, and local archives; on library websites; or, perhaps, in the Internet Archive. To complement old, paper-based records, archivists are now routinely acquiring modern, digital-only collections [Redwine and Barnard 2016]. The unique materials managed by archives and special collection libraries now include scanned paper records and 'born-digital' collections such as email, web archives, disk images, personal digital materials, and networked documents [AIMS Work Group 2012]. These materials originate with people of many backgrounds and life experiences, as well as with organizations like universities, nonprofits, businesses, and governments [O'Toole and Cox 2006].

Why is it so difficult for people to find and use archives? In part, because it is difficult for archivists to efficiently describe them. Having humans create archival descriptive information is very time and labor intensive [Yaco 2008]. Furthermore, the software that archivists use to record metadata is hobbled by user interfaces that mirror the structure of a complex, hierarchical, and rigid XML schema. This makes data entry laborious at best and error prone at worst [Prom et al. 2018]. The software used in related domains also forces people to manually create or import descriptive metadata [Bailey and Vidyarthi 2010]. In spite of the thoughtful work completed by archivists, librarians, and information technologists to acquire, preserve, organize, digitize, describe, list, and catalog such records, most of them remain under-described or squirreled away in information silos, such as local databases. Relatively little of this metadata is provided to national and international indices, much less in easily consumable formats.

While all of these factors help explain why archival metadata is so difficult to create and use, they share a "meta-problem," or a problem behind the problem. Archival theories, practices, techniques and models were developed for a pre-internet age and must be reoriented to take advantage of machine approaches [Bailey 2013]. Our field must reorient its data-sharing practices toward the approaches used in the larger internet ecosystem, as reflected in de facto standards or accepted WC3 best practices [Rubinstein 2017].

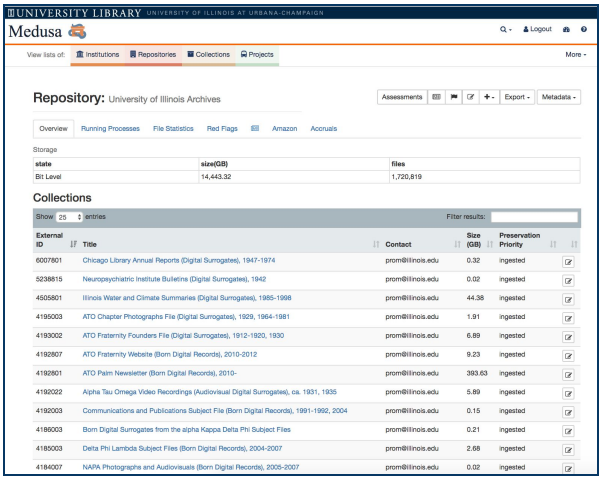
To date, the archival community has generated many conceptual analyses pointing to ways in which the description and access of historical records might be reoriented [Bak 2012, Bunn 2014, Gilliland 2014, Higgins, Hilton, and Dafis 2014, Lemieux 2014, Lemieux 2015, Owens 2014, Phillips 2012, Weisbrod 2016, Yeo 2014]. In practice, the community has begun experimenting with alternate approaches, in which machine-generated metadata augments or complements the archivist's thoughtful work [Elings 2016, Gracy, 2015, Lyons 2015]. These projects have achieved valuable local successes or apply to just one type of documentation. For example, the EPADD ("Email: Process, Appraise, Discover, and Deliver") software uses natural language processing and entity extraction tools to generate an interface for browsing and accessing email collections [Schneider and Chan 2016]. In a parallel track, the growing field of digital humanities has proposed specific ways that machine learning and natural language tools can enhance access to historical materials [Piotrowski 2012], but this work has often been applied to specialist projects.

Statement of Work and Research Implications

In this project, we propose to test the application of advanced computational methods such as natural language processing (NLP), named entity recognition (NER), continuous active learning (CAL), and neural networks (NN) in a way not previously attempted: as integral elements of a preservation repository and digital archive service. By testing and refining these tools with a large corpus of historical records and in the context of a mature preservation service, this research ultimately aims to make advanced machine learning tools available to the majority of archives, which currently cannot afford to experiment with such techniques, through a service hosted on Amazon Web Services (AWS).

Work to Date

Over the past five years, the University of Illinois Library has been building the Medusa Preservation Repository [Rimkus and Habing 2013]. Integrating major recommendations from *Preservation Metadata Implementation Strategies* (PREMIS), Medusa provides bit-level preservation services for over 103 TB of born-digital and digitized materials that are unique to the University of Illinois. The preserved materials include textual documents, scanned photographs, audio recordings, and many other types of documentation. Medusa provides a variety of preservation services such as checksum and obsolescence monitoring, and it stores three copies of all files, two locally and one in Amazon Glacier. Over the next year, the University Library will begin migrating the primary copy of all system records to AWS.

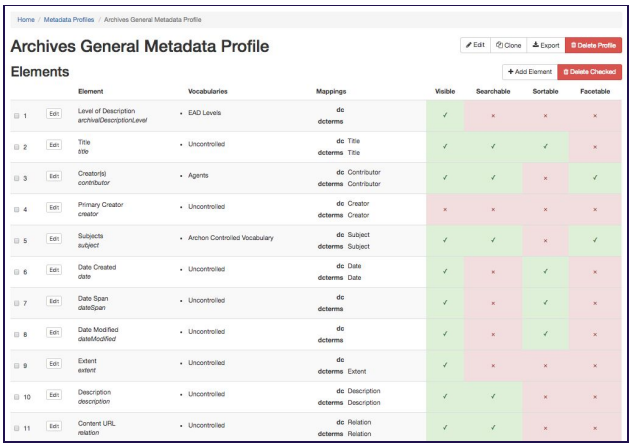


Medusa contains a large number of scanned textual documents and 'born-digital' records such as email messages, reports, and other files that contain textual information. The screenshot at right illustrates the Medusa management console and some textual records from the University of Illinois Archives. As of September 11, 2017, the text-based records deposited in Medusa comprised at least 3,320,450 files, or over 14 TB of data: a figure expected to grow exponentially over the next several years as a large volumes of text-based materials are added to the repository. As an example: On September 10, 2017, one campus office suggested that they deposit the scanned, OCR'd text from ten file cabinets into Medusa, representing just one year's worth of administrative records, shedding light on the history of the University of Illinois.

In the summer of 2015, we began developing a digital library application as a means to publish materials in the Medusa preservation repository (digital.library.illinois.edu). At first, the published collections comprised only scanned photographs, but now the digital library includes an innovative file browser application, which the public can use to navigate through large collections of textual documents. The screenshot at left shows one textual document salvaged from the disk image of a microbiologist's lab computer. The application provides a rudimentary means to access bulky collections of documents for which we have little item-level metadata. While some of these records include summary metadata notes describing the entire collection, most folders or files remain undescribed. It is not feasible for humans to create item-level descriptive metadata for all of them without assistance from a machine. Yet without such

metadata they will be found mainly through happenstance or serendipity.

The Library has taken preliminary steps to improve access to these records and to prepare the way for machine-assisted indexing. In accord with archival best practices, staff members have prepared collection-level descriptive metadata, using the recommended best practices from our community [Society of American Archivists 2013]. In addition, we are augmenting the digital library's metadata-management system, which includes methods to bulk import and to edit metadata. Specifically, we are developing a means to treat the names of people and organizations as controlled entities. That way a single person or organization might be related as a node (or vertex) along particular relationships (or edges) to multiple documents. As shown in the screenshot, we are developing a method to describe folders and items using a general metadata profile for archives. Particular documents, photographs, audio materials, and other resources might be grouped or related by named entities or provided in a time-ordered document stream.



Finally, the University of Illinois Library was recently awarded a pilot grant from the National Endowment for the Humanities to digitize and provide access to "The Cybernetics Thought Collective." [Anderson 2017]. This project is applying NLP, NER, and NN technologies to some specific correspondence files, found in the personal archives of 'cybernetics' scholars. (This field developed some of the philosophical underpinnings for modern computing.) The project provides us with direct NLP and NER development experience. We have also gained access to UIUC developer communities and a project advisory board including representatives from industry (including Stephen Wolfram, CEO of Wolfram Research). These groups are helping us to assess, implement and refine the tools with a specific set of records, and selected members of the group will be asked to continue on an advisory board for the project proposed in this document.

Taken as a whole, the work completed to date places the University of Illinois Library in an enviable position, able to undertake the next round of research and development. Prior to the proposed start date of this project (August 21, 2018), we will have the following elements in place:

- A test instance of our preservation repository running on Amazon Web Services, including direct access to selected content from the University of Illinois. (Over time, we plan to move all content into AWS, but we may not be completely done by project start date.)
- An open-source code repository on github for both preservation repository and digital library (https://github.com/medusa-project)
- An online browsing tool that provides access to large amounts of textual materials in browser
- A lightweight metadata profile and the means to automate metadata creation
- Significant local experience implementing NER, NLP, and NN technologies locally.
- The basis for experimenting with other AWS technologies that could revolutionize access to historical records (for example: Rekognition to identify people in old photos, Polly to convert text to speech and help non-sighted users access historical documents)
- Partnerships with academic computer and information science programs, as well as industry partners (Wolfram Reserach, Amazon) that have deep connections in the academic community.

Goals and Objectives

Overall, the project seeks an improved means of preserving and making available 'born-digital' records, both at the University of Illinois and elsewhere. Specifically, we propose to:1) conduct a national search resulting in the hiring of a postdoc or faculty resident with significant experience in NLP, NER, CAL, or NN technologies; and 2) direct him or her in a research project that aims to develop a replicable, AWS-based digital preservation and library application. This application will include an integral microservice that assists humans in accurately describing large, unwieldy, text-based collections. Unlike other products in

the library market, the application would integrate with and extend industry-standard technologies available through AWS. Specifically, staff will be able to edit the results of the NER, NLP, CAL, and NN technologies to classify and describe records based on previously hidden relationships, clustering, entity reconciliation, and other results from the machine processes. The end goal is not only to apply these technologies to records at the University of Illinois, but (eventually) to develop a hosted, cost recovery service, so that the digital preservation and digital library application can be marketed to and used by other libraries, archives, and museums.

Plan of Work

The incumbent postdoc/resident will work under direction of the PI and with input from a project advisory board, including members of the UIUC library, computer, and information science community; industry partners (Wolfram Research and Amazon); and external advisors. The postdoc will engage in the following tasks:

- Apply NLP, NER, NN tools (including AWS Artificial Intelligence Tools such as Deep Learning AMI's and ApacheMX Net), to extract metadata and build metadata networks from large text collections in University of Illinois Medusa repository.
- Build services and dashboard that allow archivists to run these tools on sets of records and to expose created metadata in web editing forms, for archivists to augment and improve their metadata, discovery, and search services.
- Ingest outputs into digital library services. These will run on AWS database tools (we are currently using Postgres, but may explore use of Amazon Aurora and DynamoDB with this metadata)
- Partner with other University of Illinois Library employees to refine existing digital library search and discovery services to take advantage of project outcomes.

During the project, the project PI will partner with the University's Office of Technology Management, as well as other Library faculty and staff and the project advisory board to conduct a preliminary assessment of the effectiveness of the approach taken, both in research and commercial terms. This will allow us to consider not only the specific contributions that machine learning and natural language processing make in improving access to historical materials, but also the potential feasibility of the service as a cost-recovery, hosted solution for other libraries and archives. The preliminary business assessment will include cost and revenue projections targeted at small to medium size archives. (These programs lack dedicated development staff and might benefit most directly from the proposed service.) To complete this work, the PI will pursue a project with the College of Business Illinois Business Consulting Program, the largest student-run management consulting organization in the United States (www.ibc.illinois.edu). Their report will serve as the basis for future planning with our Office of Technology Management.

Dissemination and Outcomes

- All code developed under the Amazon research award will be made available under an open source license at github.com/medusa-project
- We will present results of the research at appropriate conferences such as IEEE Big Data workshops on computational archival science (December 2018), IS&T or Archiving 2019, or the Joint Conference on Digital Libraries (JCDL). Travel will be supported by endowment funds.
- The team will submit articles to peer reviewed journals like the Journal of the American Society for Information Science and Technology (JASIST) and *College and Research Libraries.*
- Hosted service feasibility plan will be developed in consultation with Illinois Business Consulting.

Budget

Postdoc salary	Postdoc benefits @ 38.06%	Consulting Services	**Total Budget Request**	AWS Services, as described in narrative
$55,000	$20,933	$4,000	**$79,933**	$20,000 in kind contribution

Reference List

AIMS Work Group. 2012. “Born-Digital Collections: An Inter-Institutional Model for Stewardship.” https://dcs.library.virginia.edu/files/2013/02/AIMS_final_text.pdf.

Anderson, Bethany. 2017. “University of Illinois Archives Awarded NEH Grant to Digitize ‘The Cybernetics Thought Collective.’” *University of Illinois Archives*. https://goo.gl/axtr7s

Bailey, Jefferson. 2013. “Disrespect Des Fonds: Rethinking Arrangement and Description: Rethinking Arrangement and Description in Born-Digital Archives.” *Archive Journal*, 3. https://goo.gl/vgLbU5.

Bailey, Steve, and Jay Vidyarthi. 2010. “Human-Computer Interaction: The Missing Piece of the Records Management Puzzle?” *Records Management Journal* 20 (3): 279–90. https://goo.gl/ECuQVM.

Bunn, Jenny. 2014. “Questioning Autonomy: An Alternative Perspective on the Principles Which Govern Archival Description.” *Archival Science* 14 (1): 3–15. https://goo.gl/AZJTTG.

Bak, Greg. 2012. “Continuous Classification: Capturing Dynamic Relationships Among Information Resources.” *Archival Science* 12 (3): 287–318. https://goo.gl/g423Bh.

Gilliland, Anne J. 2014. “Reconceptualizing Records, The Archive, and Archival Roles and Requirements in a Networked Society.” *Knygotyra: Book Science* 63 (63): 17–34. https://goo.gl/1mwHR3.

Gracy, Karen F. 2015. “Archival Description and Linked Data: A Preliminary Study of Opportunities and Implementation Challenges.” *Archival Science* 15 (3): 239–94. https://goo.gl/9iR7JN.

Higgins, Sarah, Christopher Hilton, and Lyn Dafis. 2014. “Archives Context and Discovery: Rethinking Arrangement and Description for the Digital Age.” In Arxius i Industries Culturals. Girona, Spain. http://www.girona.cat/web/ica2014/ponents/textos/id174.pdf.

Lemieux, Victoria L. 2014. “Toward a ‘Third Order’ Archival Interface: Research Notes on Some Theoretical and Practical Implications of Visual Explorations in the Canadian Context of Financial Electronic Records.” *Archivaria* 78. https://goo.gl/HRYjwJ.

Lemieux, Victoria L. 2015. “Visual Analytics, Cognition and Archival Arrangement and Description: Studying Archivists’ Cognitive Tasks to Leverage Visual Thinking for a Sustainable Archival Future.” *Archival Science* 15 (1): 25–49. https://goo.gl/1F3tpp.

Lyons, Bertram. 2015. “Reading In: Analyzing Embedded Metadata in Digital Images.” *Medium*. July 6. https://medium.com/on-archivy/reading-in-69ab566d1e#.azumd4l6t.

O’Toole, James M., and Richard J. Cox. 2006. *Understanding Archives & Manuscripts*. Archival Fundamentals Series. Chicago, IL: Society of American Archivists.

Owens, Trevor. 2014. “The ePADD Team on Processing and Accessing Email Archives | The Signal.” Webpage. October 20. https://goo.gl/oAQiHz.

Phillips, Meg. 2012. “More Product, Less Process for Born-Digital Collections: Reflections on CurateCamp Processing.” *The Signal*. August 22. https://goo.gl/7hShNQ. .

Piotrowski, Michael. 2012. “Natural Language Processing for Historical Texts.” *Synthesis Lectures on Human Language Technologies* 5 (2): 1–157. https://goo.gl/kN2Ezo.

Prom, Christopher J., Giovanni Michetti, Katherine Timms, Andrea Tarnawsky, and Richard Pearce Moses. 2018. “Archival Arrangement and Description in the Cloud: A Preliminary Analysis.” In *Born Digital in the Cloud: Challenges and Solutions, Presentations at the 21th Archival Sciences Colloquium of the Marburg Archives School*. Preprint available at https://goo.gl/2fQj8V.

Redwine, Gabriela, and Megan Barnard. 2016. “Collecting Digital Manuscripts and Archives.” In *Appraisal and Acquisition Strategies*, 67–116. Chicago: Society of American Archivists.

Rimkus, Kyle R., and Thomas Habing. 2013. “Medusa at the University of Illinois at Urbana-Champaign: A Digital Preservation Service Based on PREMIS.” In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, 49–52. Indianapolis, IN: ACM Press. https://goo.gl/dHdZRn.

Schneider, Josh, and Peter Chan. 2016. “Let the Entities Describe Themselves.” *bloggERS!* May 3. https://goo.gl/Yx6Hia.

Society of American Archivists. 2013. *Describing Archives: A Content Standard*. 2nd edition. Chicago, IL: Society of American Archivists. https://goo.gl/91S8zj.

Weisbrod, Dirk. 2016. “Cloud-Supported Preservation of Digital Papers: A Solution for Special Collections?” *LIBER Quarterly* 25 (3). https://goo.gl/4upvny.

Yaco, Sonia. 2008. “It’s Complicated: Barriers to EAD Implementation.” *The American Archivist* 71 (2): 456–75. https://doi.org/10.17723/aarc.71.2.p0h0t68547385507.