Abstract

Developing a Computational Framework for Library and Archival Education

The University of Maryland seeks $99K to hold a major planning Workshop of library and archival educators and technologists, through a one-year Laura Bush 21st Century Librarian (LB21) National Forum Grant under the National Digital Platform funding priority. This will help develop the building blocks of an integrated library and archival Master's level educational curriculum to educate the next generation of librarians and archivists in the computational treatments of collections. The Workshop will be held in conjunction with the 2019 iConference meeting hosted at the University of Maryland in March / April 2019.

The last two years have seen the emergence of the concept of "Collections as Data" in cultural heritage institutions, where computational methods and tools are increasingly leveraged to enhance library and archives collections. The U. Maryland iSchool is an active member of this community and has furthered the research and curricula development with computational treatments of collections through developments in Computational Archival Science (CAS) in particular. The U. Maryland project team comprises computer scientists, software developers, archivists, and educators, now assembled at the University of Maryland iSchool and collaborating in the Digital Curation Innovation Center (DCIC).

In the first six months of this project the current core knowledge of Library and Archival Science professionals will be identified through review of current MLIS program curricula and review of guidance from Library Science and Archival Science professional, and program certification associations. This will be carried out with student participation and feedback. Next, we will focus on computational concepts, methods and practices that are being taught in MLIS programs and this knowledge represented as units. The current Library and Archival Science research and practice literature will be reviewed to identify computational concepts, methods and practices that complement the Library and Archival Science concepts, methods and practices already included in the Library and Archival Science knowledge base. We anticipate that one of the key computational units will be "Computational Thinking". We call this knowledge base *"The Computational Framework for Library and Archival Education"*. We will show how Library and Archival Science curricula can be developed from this knowledge base that include core Library and Archival Science topics as well as complementary computational topics.

For the Workshop, we plan on inviting iConference attendees actively contributing to the "Collections as Data" field, US thinkers, a small set of international participants, and prominent Washington D.C. computational and library / archives collection leaders.

We will disseminate the products of this project at the annual ALA and SAA professional library and archives meetings. One of the objectives is to seek acceptance of professional associations of the framework for consideration as the body of knowledge from which curricula can be developed. The goal is to create a community of iSchool Educators and Researchers to implement curricula that incorporate Computational Thinking and computational concepts, methods and practices into their MLIS programs.

We expect this will lead to the submission of a major grant proposal to implement and refine the designed framework and computational curriculum. This would be tested at UMD's iSchool in collaboration with other participating iSchools, with the goal of creating a shared, virtual, reusable, and modular educational resource for the library and archival education community at large. Our approach includes identifying canonical collections, analytical methodologies, and open source software applications, and developing an initial set of end-to-end repurposable lesson plans to be disseminated and shared with the community at large. The goal is to create a sustainable community of iSchool Educators and Researchers to implement curricula that incorporate Computational Thinking and computational concepts, methods and practices into their MLIS programs.

Developing a Computational Framework for Library and Archival Education

The University of Maryland seeks $99K to hold a major planning Workshop of library and archival educators and technologists. This Workshop will identify the foundation / building blocks for an integrated library and archival Master's educational curriculum centered on computational treatments of large complex collections. This is designed to prepare the next generation of librarians and archivists to meet the evolving needs of professionals working with digital collections. This Workshop will be held in conjunction with the 2019 iConference meeting hosted at the University of Maryland in March / April 2019.

1. Statement of Broad Need

The last two years have seen the emergence of the concept of "Collections as Data" in cultural heritage institutions, where computational methods and tools are increasingly leveraged to enhance library and archives collections:

> *"Combined with an increasing flow of born-digital items, digital library collections have come to represent a rich community resource for users… Yet a focus on replicating traditional ways of interacting with collections in a digital space does not meet the needs of the researcher, the student, the journalist, and others who would like to leverage computational methods and tools to treat digital library collections as data."*[1]

In the "*Always Already Computational: Library Collections as Data"* IMLS-funded project, Thomas Padilla illustrates several typical "computational" treatments of collections where "a Digital Humanities researcher engages in term frequency visualization, topic modeling, and network analysis across thousands and sometimes even millions of items." He adds, that in addition to the computational treatment of text data "the scope of data extends to images, moving images, sound, web archives, and beyond."

Two years ago, a new peer-reviewer journal was launched, the ACM Journal on Computing and Cultural Heritage (JOCCH). This journal brings together an interdisciplinary community focusing on digital infrastructure for cultural heritage projects where big data technological stacks and graph databases are increasingly used to manage large distributed collections, with computational techniques spanning: genetic algorithms, virtual 3D, image segmentation, automated classification, and visual analytics.[2]

The U. Maryland iSchool is an active member of this community and has furthered the research and curricula development with computational treatments of collections[3]. This can be seen through their work on Computational Archival Science (CAS)[4], defined as:

> *An interdisciplinary field concerned with the application of computational methods and resources to large-scale records /archives processing, analysis, storage, long-term preservation, and access, with the aim of improving efficiency, productivity and precision in support of appraisal, arrangement and description, preservation and access decisions, and engaging and undertaking research with archival materials.*

In this work, CAS treatment of collections are being explored through eight case studies which include: (1) evolutionary prototyping and computational linguistics; (2) graph analytics, digital humanities, and

---

[1] Padilla, T, et al. (2016): *"Always Already Computational: Library Collections as Data"*. Link: https://www.imls.gov/grants/awarded/lg-73-16-0096-16

[2] ACM Journal on Computing and Cultural Heritage (JOCCH): https://jocch.acm.org/index.cfm

[3] Marciano, R., et al. (2018): *"Archival records and training in the Age of Big Data."* In J. Percell , L. C. Sarin , P. T. Jaeger , J. C. Bertot (Eds.), Re-Envisioning the MLS: Perspectives on the Future of Library and Information Science Education (Advances in Librarianship, Volume 44B, pp.179-199). Emerald Publishing Limited. Link: http://dcicblog.umd.edu/cas/wp-content/uploads/sites/13/2017/06/Marciano-et-al-Archival-Records-and-Training-in-the-Age-of-Big-Data-final.pdf.

[4] The "Computational Archival Science (CAS) Portal". Link: http://dcicblog.umd.edu/cas/

archival representation; (3) computational finding aids; (4) digital curation; (5) public engagement with (archival) content; (6) authenticity; (7) confluences between archival theory and computational methods: cyberinfrastructure and the records continuum; and (8) spatial and temporal analytics. In addition, each of these case studies concludes with a "Takeaways for CAS/MLS Education" statement.

Also in the last two years, the Library of Congress (LOC) has launched a new group called the National Digital Initiatives, whose goals are to leverage computation to develop new knowledge from collections, and enable interactions on digital platforms. Two national "Collections as Data" events have already taken place:

> *"The rise of accessible digital collections coupled with the development of tools for processing and analyzing data has enabled researchers to create new models of scholarship and inquiry,"*[5] *and "More relevant, more accessible, more visual, and more useful--these are some benefits of making digital collections available as data and ready for computational analysis."*[6]

At the 2018 International Digital Curation Conference (IDCC2018) in Barcelona, Spain, keynote speaker Luis Martínez-Uribe discussed the need to blend analytics and digital curation. Examples included curated data using supervised learning for categorization, clustering methods that support the process of entity disambiguation, using off-the-shelf AI for automatic transcription of audio, and sentiment analysis / keyword extraction from text. His proposal is to embed computational approaches into the curation lifecycle itself to change the way in which practicing archivists and librarians gain insights into collections while they are actively being processed.[7]

Finally, the Smithsonian Institution has conducted cutting-edge work showing how computational treatments using artificial intelligence can revolutionize archival museum research.[8] Deep learning software is used to help botanists with plant categorization at museums with over 5 million scanned specimens. Two big data analytics questions are pursued: (1) With what accuracy can a trained neural network sort mercury-stained plan specimens from clean ones? and (2) With what accuracy can machine learning algorithms recognize members of two similar plant families? Preliminary results are outstanding.

There is a well-known and critical need to extend iSchool curricula into this space to better prepare graduate students to manage digital collections in order to address future needs of these new types of researchers and workers. The IMLS has funded many of the "Collections as Data" related projects. In "NDP at Three"[9] the skills gap with practicing librarians and archivists and the management gap with institutional administrators are prominently addressed. We intend to address the crucial training gap with iSchools by convening the principal leaders of these IMLS-funded initiatives and other key professionals (see Project Design) in the context of a major international meeting. We seek a Laura Bush 21st Century Librarian (LB21) National Forum Grant under the National Digital Platform funding priority to develop the building blocks of an integrated library and archival Master's level educational curriculum to educate the next generation of librarians and archivists in the computational treatments of collections. Only through adequate preparation of this next generation of practitioners, can the potential for "Collections as Data" be unlocked, and the culture of libraries and archives extended into the realm of computational treatments and processing of collections at scale.

---

[5] Collections as Data: Stewardship and Use Models to Enhance Access (Sep. 27, 2016). Link: http://digitalpreservation.gov/meetings/dcs16.html

[6] Collections as Data: IMPACT (Jul. 25, 2017). Link: http://digitalpreservation.gov/meetings/asdata/impact.html

[7] "Blending Analytics and Curation: Data Explorations from a Library in a Cultural Organization", Luis Martínez-Uribe: at IDCC2018, Feb. 21, 2018. Link: http://www.dcc.ac.uk/sites/default/files/documents/IDCC18/PresentationsIDCC18/LMUribe_BlendingCurationAnalytics_IDCC2018.pdf

[8] Smithsonian Magazine, Nov. 3, 2017. Link: https://www.smithsonianmag.com/smithsonian-institution/how-artificial-intelligence-could-revolutionize-museum-research-180967065/

[9] NDP at Three (The First Three Years of IMLS Investments to Enhance Digital Platform for Libraries), Sep. 2017. Link: https://www.imls.gov/sites/default/files/publications/documents/imls-ndp-three-508.pdf

2. Project Design

Table 1 in the supplemental *Supportingdocument1.pdf* attachment shows a sample of the Project Team's recent research and curricula development innovations, through funded projects and novel training programs, with a focus on "Collections as Data". For the last two years, the UMD iSchool Digital Curation Innovation Center (DCIC)[10], whose mission is to:

> *"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"*

has pursued a "Computational Archival Science" training and teaching agenda. We have created innovative classes that blend elements of archival and computational thinking, in project-oriented hands-on team-based environments[11], including the launch of a first graduate seminar in computational archival science in fall 2017. Our proposed 12-month Workshop grant is meant to help design a draft Library and Archival Educational Curriculum centered on computational treatments of large complex collections.

2.1 Pre-Workshop Planning

In the first six months of this project (October through March), leading up to the March / April iConference Workshop event, we are planning a number of synergistic activities. Two new permanent courses were developed by the Digital Curation Innovation Center (DCIC) faculty and staff and will first be offered in the Fall 2018 to graduate students in the MLIS Master's Program. This will provide them the opportunity to integrate their classroom education with technology through hands-on research experience on DCIC projects related to "Collections as Data". Highlights include:

1. *INST 747: Research in Advanced Digital Curation:* Class/team projects will focus around six major themes: community displacement, movement of people, citizen internment, racial zoning, the legacy of slavery, and cyber-infrastructure development. Project participants will work with external stakeholders, and gain experience in applying digital curation tools, methods, and analytics to large, complex archival collections that form the basis for the projects.
2. *LBSC 789A: Special Topics in Contemporary Archives:* This seminar will offer students the challenging opportunity of collaborating with DCIC faculty and staff, along with external archival experts, to design a model archival curriculum integrated with computational methods and tools needed for 21st century archival practice.

These experimental and innovative classes will be co-taught by faculty (Richard Marciano, Ken Heger, Bill Underwood, and Michael Kurtz), and research staff / former students (Mary Kendig). The goal is to engage MLIS students in co-designing an initial set of "Collections as Data" learning modules which will be presented at a special curricula development session at the 3rd CAS Workshop at the IEEE Big Data Conference 2018[12] in Seattle. The IEEE Big Data 2017 conference had close to 1000 participants from 50 countries. The initial set of learning modules will frame the discussions at the March/April 2019 Workshop at the iConference event at the University of Maryland.

Mary Kendig, former MLIS and MIM graduate student at the UMD iSchool, who helped design DCIC computational projects, will provide key leadership in working with students. In her Library of Congress

---

10 Digital Curation Innovation Center (DCIC): http://dcic.umd.edu

11 Marciano, et al.: "*Practical Digital Curation Skills for Archivists in the 21st Century*", presented at MARAC 2016: https://drum.lib.umd.edu/handle/1903/18865

12 IEEE BigData 2018 3rd CAS Workshop. Link: http://dcicblog.umd.edu/cas/ieee-big-data-2018-3rd-cas-workshop/

blog[13], she discussed the critical need to: encourage students to enroll into technology intensive courses and programs, develop projects for students to gain digital skills and hands-on experience, and develop collaborations with institutions to provide beneficial learning environments for students. Other students in the DCIC have also reflected on ways of "Unlocking the Science of the Past", "Developing Computational Finding Aids", and using "Network Analytics and Graph Databases" with big cultural content[14].

In the new seminar INST747, students will get credit while participating in hands-on collaborative group projects in the DCIC. These have included: (1) Justice, Human Rights & Cultural Heritage projects (related to community displacement, migration, citizen internment, refugee narrative, racial zoning, and the legacy of slavery), and (2) Cyberinfrastructure for the curation & management of digital assets at scale (DRAS-TIC Fedora, Brown Dog, and Hands-on computation in the archives). We have acquired precious feedback and gained deep insights as to what appears to be the most effective ways of learning and how to distill and package these lessons and wish to take this knowledge to the next level. In the 2017 / 2018 academic year, over 100 students volunteered in the DCIC, and they can now gain class credit.

In the seminar LBSC789A, the current core knowledge of Library and Archival Science professionals will be identified through review of current MLIS program curricula and review of guidance from Library Science and Archival Science professional and program certification associations. This knowledge will be represented as knowledge areas and units within those areas. Additional non-computational units taught in MLIS courses will be added to the core knowledge. Next, the seminar will focus on computational concepts, methods and practices that are being taught in MLIS programs and this knowledge represented as units. Then the current Library and Archival Science research and practice literature will be reviewed to identify computational concepts, methods and practices that complement the Library and Archival Science concepts, methods and practices already included in the library and Archival Science knowledge base. We anticipate that one of the key computational units will be "Computational Thinking". Computational thinking is considered a form of problem solving that uses modeling, decomposition, pattern recognition, abstraction, algorithm design, and scale[15]. All units will be refined into topics and learning objectives. References will be provided to detailed knowledge sources. We call this knowledge base *"The Computational Framework for Library and Archival Education"*. We will show how Library and Archival Science curricula can be developed from this knowledge base that include core Library and Archival Science topics as well as complementary computational topics. Table 2 shows some of the lessons learned so far and computational thinking mapping to archival concepts so far:

Archival Concepts	Computational Methods
Support accessibility to large historical European Commission archival holdings	Topic Modeling for concept extraction from large EC archival holdings
Going from paper catalog entries to digital catalogs, Matching records in distributed databases.	Graph and Probabilistic Databases
Technology assisted review accessibility of presidential and federal e-mail accessioned into National Archives	Analytics, predictive coding to address PII
Provenance of scientific data records (datasets). Trust in authenticity of the data, transparency and reuse	DataONE extensions to PROV (Provenance data model)

13 "The Reality of Preparing MLS Students for a Competitive and Increasingly Digital World", Guest blog post by MLIS graduate student, Mary Kendig in The Signal, Library of Congress: Apr. 14, 2017. Link: https://blogs.loc.gov/thesignal/2017/04/identity-crisis-the-reality-of-preparing-mls-students-for-a-competitive-and-increasingly-digital-world/

14 Education & Training blogs contributed by DCIC MLIS graduate students (Jennifer Proctor and Claire McDonald), Fall 2017. Link: http://dcicblog.umd.edu/cas/#Education&Training

15 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

Need for a service to assign globally unique persistent identifiers to data sets in order to support accessibility, reference and reuse.	Scalable, robust automated computational service for data content comparison.
Enriched Archival Science concepts	Linguistic Models and Graph Theory
Provenance in terms of why, who and how	Abstraction and ontology construction
Web Archives Research Objects — Disciplinary perspective, legal agreements, Motivations, Interpretation, Designs	Research Objects Framework used to analyze the computational methods used in web archives research, — Research Objects in Computational Science
Appraisal	File Format Characterization, File Format policies, Bulk extractor (Identifies PII), Content Preview, Tagging
Corpus — One Billion Requests for Linguistic Services	Text mining
Trusted digital repositories (TDR), OCR, cultural heritage platforms	EUDAT automated scalable e-infrastructure, integrated computation services,
Support accessibility to large historical European Commission archival holdings	Topic Modeling for concept extraction from large EC archival holdings
Support accessibility to large historical European Commission archival holdings	Topic Modeling for concept extraction from large EC archival holdings
Annotation, entity extraction, NLP, machine learning	Archival materials contextual discovery
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
Classification of archival images	Line detection, image segmentation
Recordkeeping	Auto-categorization, auto-classification, e-discovery, machine learning
Iterative design, value-sensitive design	Heuristics for CAS research,
Knowledge complexity in archives	Digital narrative with big data,
Personally Identifiable Information (PII)	NLP, NER, sentiment analysis
Classification of time-coded collections of textual collections into epochs and periods	Cultural analytics, topic modeling
Structured data interfaces to archival materials	APIs for cultural heritage materials, graph databases
Decentralized recordkeeping	Blockchain, secure computing, trustworthiness
Recordkeeping, digital preservation, archival trust	Blockchain, computational validation, distributed ledger, computational trust

Table 2: Lessons learned at the 1st two CAS Workshops

While this mapping based on earlier CAS Workshops provides valuable insights, a more systematic study of the existing body of knowledge needs to be pursued in the first six months of the project.

Out of this body of knowledge a course syllabus on Computational thinking will be developed in the first 6 months of this project, where library and archives students are taught problem solving that uses modeling, decomposition, pattern recognition, abstraction, algorithm design, and scale[16]. All units will be refined into topics and learning objectives

We will work closely with Dr. David Weintrop who studies the design of computational learning environments. His interests are at the intersection of the Learning Sciences, human-computer interaction, computer science education, with an emphasis on the application of theory to the design and evaluation of innovative computer science education. His earlier work has focused on advancing the design of environments and technologies that can broaden participation in computing, including bringing learners from

16 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

divers and historically underrepresented backgrounds into computing.[17]

2.2 iConference Workshop

The DCIC seeks to build on the "Collections as Data" work in progress by leveraging a significant international conference, the 2019 iConference, which will be hosted in the U.S. at the University of Maryland (after taking place in China and the UK). This conference has grown to 600 participants in the last couple of years and engaged over 200 universities in 33 countries with representation from Asia, Africa, North America, the Caribbean, South America, Western Europe, Eastern Europe, and the Middle East. It has provided a forum for interdisciplinary professors and administrators working in iSchools and other departments, to share research on where technology and information science are heading. The iConference follows the iSchool spectrum, with representation from fields such as: Libraries, Archives, Digital Curation, and others including Human-Computer Interaction, Computer-Supported Cooperative Work, Information Science, Analytics, and Information Infrastructure, and Learning Sciences and Education.

The Computational Framework for Library and Archival Education will be presented at the two-day "Collections as Data" Workshop at the U. of Maryland, which we will host at the end of March 2019 in conjunction with the 2019 iConference[18]. The draft syllabus for the course on Computational Thinking for library and archival students will also be presented. The invited participants will share their insights and experiences, and also help critique and refine the curriculum. Finally, we will publish and disseminate the vetted findings.

We plan on inviting four categories of attendees to the Workshop:

I. **iConference attendees** actively contributing to the Collections as Data space, such as: (a) **Jane Greenberg**, professor from *Drexel University's iSchool*, with her IMLS-funded Library Education and Data Science – LEADS project for doctoral training, (b) **Katie Shilton**, Associate Professor at the UMD iSchool, with her the NSF-funded PERVADE project on big data ethics, (c) **Greg Jansen,** Principal Software Engineer at the UMD iSchool DCIC Center, and co-PI on the IMLS-funded DRAS-TIC Fedora project on collection infrastructure scalability.

II. **Other key US thinkers** in the emerging field of Collections as Data, including: (a) **Thomas Padilla**, Visiting Digital Research Services Librarian at the *University of Nevada Las Vegas*, on the seminal IMLS-funded project "Always Already Computational: Library Collections as Data", (b) **John Chodacki**, Director of the *University of California Curation Center at the California Digital Library*, on the IMLS-funded "Expanding Library Carpentry" project, (c) **Chris Erdmann,** Library Carpentry Community and Development Director, formerly at *NCSU Libraries* part of an IMLS-funded grant headed by researchers at the University of Pittsburgh, called "The Data Scientist as the 21st Century Librarian?," and (d) **Ceilyn Boyd**, Research Data Program Manager at the *Harvard Library*, **Jessica Farrell**, Curator of Digital Collections at the *Harvard Law Library*, who recently hosted an unconference on Computational Archival Science in December 2017.

III. **A small set of international participants** with unique expertise and knowledge that would be beneficial to the conversation and help create positive impact for US libraries: (a) **Mark Hedges,** Senior Lecturer in the Department of Digital Humanities at *King's College London* – Chair of the last two Computational Archival Science workshops and the last three Big Humanities Data workshops all part of the IEEE Big Data Conference series, (b) **Victoria Lemieux**, Associate Professor in

[17] Computational Thinking in STEM, by David Weintrop. See: http://www.terpconnect.umd.edu/~weintrop/#projects?ct-stem

[18] The iConference has grown to 600 participants from 200 universities in 33 countries. It provides an interdisciplinary forum to share research on where technology and information science are heading. See: http://ischools.org/the-iconference/about-the-iconference/iconference-preview/

Archival Science at the *University of British Columbia iSchool, Canada* – Editor of "Building Trust in Information" and Cluster Lead of the Blockchain@UBC Research Cluster, where she is an expert in provenance and blockchain technology for trustworthy digital repositories, and (c) **Yoichi Tomiura**, Professor at *Kyushu University, Japan* – Deputy Director General of University Libraries and expert in computational linguistics in the Library Science and Informatics department.

IV. **Prominent Washington D.C. computational and library/archives collection leaders**: **(a) Kate Zwaard**, Director of Digital Strategy at the *Library of Congress,* and founder of the National Digital Initiatives "Collections as Data" group, **(b) Paul Wester**, Director of the U.S. Department of Agriculture's *National Agricultural Library* in Beltsville, MD, **(c)** experts from the *Smithsonian Institution*, including **Rebecca Dikow**, Research Data Scientist in the Data Science Lab at the Office of the CIO, and **Bob Horton**, Assistant Director of Collections and Archives at the National Museum of American History, **(d)** large record collections, access, and PII experts from the *National Archives and Records Administration (NARA)*, **TBD**, and **(e)** Computational History experts from the *U.S. Holocaust Memorial Museum (USHMM)*, including **Michael Levy**, Director of Digital Assets Management and Preservation & **Michael Haley Goldman**, Director of Future Projects.

These invitees cover a very large swath of expertise, including ethics, policy, and privacy, and are all engaged in educational and learning activities.

During Day 1 of the Workshop, the "Computational Framework" will be presented to the participants. They will be shown a "Computational Archival Science" curriculum developed from the "Computational Framework". The curriculum will be illustrated with course syllabi or learning modules including the Syllabus for Computational Thinking for Librarians and Archival Scientists. Participants will share their perspectives and experiences on Collections as Data.

During Day 2 of the Workshop, participants will be broken into workgroups and asked to review and refine the units, topics and learning objectives, reacting to the syllabus and sharing their critical review of what the promising topics are and where the gaps remain.

The goal is to create a community of iSchool Educators and Researchers willing to implement curricula that incorporate Computational Thinking and computational concepts, methods and practices into their MLIS programs.

2.3 Post-Workshop Follow-up

Following the Workshop, the Maryland Education Digital Infrastructures (MEDI) Lab will help produce the following deliverables:

- A Computational Framework document that describes the body of knowledge from which curricula and lesson plans can be developed and/or existing curricula can be compared, including:
  - Facets and strands
  - Skills-based outcome statements
  - Pre and co-requisite learning definitions
- Open access tools for institutions that wish to deliver similar educational programs. The infrastructure will combine both the computational tools used in cloud-based student technology environments.[19] Exemplar datasets and instructions that can be used by educators and students to exercise computational archival science.
- Community development plan that includes a collaboration website to be shared by institutions that will participate in the further development of the Computational Framework for Library and Archival Education movement. We will publish the framework as a white paper on this website.

[19] Cloud-based Student Technology Workspaces: http://medi.umd.edu/node/15

- Guidance to field on risks of using computational tools with sensitive and cultural collections, including discussions of privacy and ethical use of sensitive information.

These products will be developed by the MEDI Lab using state-of-the-art cloud-based digital tools (ex: Git, Wiki), so that those inside the project and other stakeholders can provide input on the process.

We will disseminate the products of this project at the annual professional archives and library meetings where we hope to hold discussion panels: ALA Annual Conference in Washington, D.C., June 20-25, 2019, and SAA Annual meeting in Austin, TX, July 29, 2019 – August 3, 2019. One of the objectives is to seek acceptance of professional associations of the framework for consideration as the body of knowledge from which curricula can be developed. A framing white paper will also be prepared for later submission and presentation at the IEEE Big Data 2019 Conference in December, 2019.

Finally, a Virtual Seminar will be held online in September, 2019, to bring participants back together and disseminate the final results.

2.4 Project Team

The U. Maryland project team consists of collection, computational, and educational experts with a track record on developing innovative curricula and conducting research in "Collections as Data":

- **Dr. Richard Marciano** is a professor in the College of Information Studies at the University of Maryland and director of the Digital Curation Innovation Center (DCIC). Prior to that, he was a Professor at the School of Information and Library Science at the University of North Carolina at Chapel Hill for 6 years. He also conducted research at the San Diego Supercomputer Center (SDSC) at the University of California San Diego (UCSD) for 13 years. His research interests center on digital preservation, sustainable archives, cyberinfrastructure, and big data. He is also the 2017 recipient of Emmett Leahy Award for innovation in records and information management.
- **Dr. William Underwood** is an Affiliate Professor with the College of Information Studies and at the University of Maryland. He earned a doctorate in Computer Science from the University of Maryland in 1980 and was formerly on the academic and research faculty of the Georgia Institute of Technology. His current research interests are in developing formal, theoretical foundations for records management and archival science, experimental investigations of alternative digital preservation strategies, and the application of natural language processing, machine learning and knowledge-based reasoning technologies to the support of automated archival description, Freedom of Information Act (FOIA) review, and search and retrieval of records in digital archives.
- **Dr. Michael J. Kurtz** was the Associate Director of the Digital Curation Innovation Center in the College of Information Studies at the University of Maryland (up until June 2018), and will continue to partner as a consultant. Prior to this he worked at the U.S. National Archives and Records Administration for 37 years as a professional archivist, manager, and senior executive, retiring as Assistant Archivist in 2011. He received his doctoral degree in European History from Georgetown University in Washington, D.C. Dr. Kurtz has published extensively in the fields of American history and archival management.
- **Dr. Katrina Fenlon** is joining the UMD iSchool in Aug. 2018 as an Assistant Professor. Coming from the iSchool at the University of Illinois Urbana-Champaign's iSchool, Katrina is a leading expert in the investigation of digital humanities research collections and their uses and users. Through typological analysis and content analysis of a large set of cultural collections, her dissertation research characterized collections in terms of differences in their underlying data models, and identified implications of those differences for collection use, sustainability, and preservation. She will be an invaluable collaborator in exploring how to model cultural collections for computational use.

- **Dr. Adam Kriesberg** is a Lecturer at the UMD iSchool. His research and teaching focus on digital curation, research data management, and archives. He has worked as a university cooperator with the USDA National Agricultural Library on projects related to digital curation at that institution. A 2016-2017 Research Data Alliance Data Share Fellow, his work has appeared in publications including *Archival Science, The American Archivist, The International Journal of Digital Curation,* and *Data Science Journal*.
- **Dr. Philip Piety** is a Senior Lecturer and Learning Scientist at the iSchool. He is an expert in learning technologies and analytics and founder of the Maryland Education Digital Infrastructures (MEDI) lab that explores the tools that mediate pedagogy. His view of these tools is deeply sociotechnical and his book *Accessing the Education Data Movement* in the Technology Education Connection series from Teachers College Press. His current book project it titled *Social at Scale* and explores how learning technologies can be used to create deeply social learning experiences for larger numbers of students. He will be an integral part of the team that designs and develops the technology used in this project.
- **Dr. David Weintrop** is an Assistant Professor at UMD with a joint appointment in the Teaching & Learning, Policy & Leadership (TLPL) Department in the College of Education and the iSchool. Within TLPL, he is a member of the Technology, Learning, and Leadership specialization. His research seeks to understand how best to support learners in developing meaningful understandings of computational ideas and positive attitudes towards computing.
- **Greg Jansen** is the DCIC Center's Principal Software Engineer and builds data repositories with new capabilities for computation and analysis. His systems support human rights and the public interest, through archives-based historical demography, open-access government records, and cultural preservation. He has a focus on creating traceable data curation workflows that connect new data or holdings to evidence, through open workflows and chains of provenance.
- **Will Thomas** is a doctoral student in the DCIC and has been a part of the digitization of the Civil Rights Department records, assessing several hundred folders of materials and providing detailed folder-level descriptions, under the aegis of a joint initiative between the University of Maryland Libraries Special Collections and the African American Digital Humanities Initiative (AADHum). He has also been involved in work to re-contextualize this collection for broader academic and cultural audiences.
- **Mary Kendig** is a Research Administrator managing multiple grants exploring the re-use of online public data and data ethics. Prior to this, she managed research at the Digital Curation Innovation Center (DCIC) and worked at the U.S. National Archives and Records Administration for two years. She is the Conference Coordinator for the upcoming iConference in 2019.

The Project Team collectively embodies deep expertise in libraries, archives, digital / data curation, digital humanities, cyberinfrastructure, educational, and computational techniques from leading research, teaching, and records management institutions including: the San Diego Supercomputer Center, the University of Illinois Urbana-Champaign, the Georgia Institute of Technology, the University of Michigan, Northwestern University, the University of North Carolina at Chapel Hill, and the University of Maryland, and the National Archives and Records Administration. The Team comprises computer scientists, software developers, archivists, and educators, now assembled at the University of Maryland iSchool and collaborating in the DCIC. This body of pre-existing work supports our proposal's success and future outcomes.

3. Diversity Plan

The DCIC currently works with historically significant collections related to social justice, human rights, and

cultural heritage themes, including: community displacement, racial zoning, refugee narratives, citizen internment, and the legacy of slavery. In the context of exploring Collections as Data we are developing best practices for working ethically with the underlying communities those collections document as well as considering value-sensitive design. This reflects the diverse student body working with us on these projects.

Doctoral student Will Thomas is both a member of the DCIC and the African American History, Culture and Digital Humanities initiative (AADHum – http://aadhum.umd.edu). Located at the University of Maryland in the same building as the DCIC and co-directed by MITH, UMD's digital humanities center, this Mellon Foundation funded initiative seeks to prepare the next generation of digital humanists and African Americanist scholars by broadening the conversation around new theories, methods and tools to explore African American art, labor and migration. It has featured a Scholars speaker series, workshops, and a digital incubator program, which complements our computational archives and libraries endeavors. We will work with leaders of this initiative to reach out to key individuals in the field, including leaders at the National Museum of African American History and Culture (NMAAHC).

Given the geographic home base of the Project Team members, which includes Washington D.C. / Maryland, Atlanta, and Durham N.C., we have already reached out to a number of historically black colleges and universities to broaden participation at the Workshop but also to establish partnerships and present our findings in the 6-month period after the Workshop.

4. Broad Impact

Our project will produce a *"Computational Framework for Library and Archival Education"* based on the study of the current core knowledge of Library and Archival Science. This knowledge, represented as knowledge areas and units within those areas, will help identify computational concepts, methods and practices that complement Library and Archival Science concepts. We anticipate that one of the key computational units will be on "Computational Thinking". The intended impact is to create a community of iSchool Educators and Researchers willing to implement curricula that incorporate Computational Thinking into their MLIS programs. Our main contribution is to create a common vocabulary and alignment of ideas that will allow "Collections as Data" to penetrate MLIS programs and training and create a new discipline.

The Workshop will help identify the building blocks of an integrated library and archival educational curriculum centered on computational treatments of large complex collections. We expect this will lead to the submission of a grant proposal to implement and refine the designed framework and computational curriculum. This would be tested at UMD's iSchool in collaboration with other participating iSchools, with the goal of creating a shared, virtual, reusable, and modular educational resource for the library and archival education community at large. Our approach includes identifying canonical collections, analytical methodologies, and open source software applications, and developing an initial set of end-to-end repurposable lesson plans to be disseminated and shared with the community at large. The course in computational thinking for library and archival students will engender the problem solving skills needed to address the issues arising from the processing of large digital collections. This project will harmonize and lead the way in which librarians and archivists can be effectively trained to meet the challenges of the future and to help advance the needs of researchers and users wishing to benefit from emerging "Collections as Data" opportunities.

Schedule of Completion (1-year grant)

10/2018-9/2019	**2018**			**2019**								
	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep
Pre-Symposium Planning												
Experimental iSchool Classes (INST747 & LBSC789A)												
Preparation of white paper for CAS#3 conference												
Computational thinking for Library & Archival Science												
Creating initial Computational Framework												
iConference 2019 Symposium												
Presenting of Computational Framework to participants												
Refinement by participants												
Presentation of computational thinking syllabus												
Post-Symposium Planning												
Working with MEDI Lab on community development plan												
Dissemination at ALA												
Dissemination at SAA												
Preparation of white paper for CAS#4 conference												
Hosting of a online Virtual Seminar												

DIGITAL PRODUCT FORM

Introduction
The Institute of Museum and Library Services (IMLS) is committed to expanding public access to federally funded digital products (i.e., digital content, resources, assets, software, and datasets). The products you create with IMLS funding require careful stewardship to protect and enhance their value, and they should be freely and readily available for use and re-use by libraries, archives, museums, and the public. However, applying these principles to the development and management of digital products can be challenging. Because technology is dynamic and because we do not want to inhibit innovation, we do not want to prescribe set standards and practices that could become quickly outdated. Instead, we ask that you answer questions that address specific aspects of creating and managing digital products. Like all components of your IMLS application, your answers will be used by IMLS staff and by expert peer reviewers to evaluate your application, and they will be important in determining whether your project will be funded.

Instructions

- ☐ Please check here if you have reviewed Parts I, II, III, and IV below and you have determined that your proposal does NOT involve the creation of digital products (i.e., digital content, resources, assets, software, or datasets). You must still submit this Digital Product Form with your proposal even if you check this box, because this Digital Product Form is a Required Document.

If you ARE creating digital products, you must provide answers to the questions in Part I. In addition, you must also complete at least one of the subsequent sections. If you intend to create or collect digital content, resources, or assets, complete Part II. If you intend to develop software, complete Part III. If you intend to create a dataset, complete Part IV.

Part I: Intellectual Property Rights and Permissions

A.1 What will be the intellectual property status of the digital products (content, resources, assets, software, or datasets) you intend to create? Who will hold the copyright(s)? How will you explain property rights and permissions to potential users (for example, by assigning a non-restrictive license such as BSD, GNU, MIT, or Creative Commons to the product)? Explain and justify your licensing selections.

This project will produce documentation that will be freely available via a public website hosted at the UMD iSchool.

A.2 What ownership rights will your organization assert over the new digital products and what conditions will you impose on access and use? Explain and justify any terms of access and conditions of use and detail how you will notify potential users about relevant terms or conditions.

The University of Maryland will not assert ownership rights over digital products produced in the course of this project, to the extent legally permissible and subject to any obligations to third parties. These products will include a website and a collection of documents related to curricula development. These products will be free available for access and use.

A.3 If you will create any products that may involve privacy concerns, require obtaining permissions or rights, or raise any cultural sensitivities, describe the issues and how you plan to address them.

The information we plan on creating will be publicly available and does not involve human subjects. If any documents are found later to raise privacy concerns, we will immediately remove them from the project.

Part II: Projects Creating or Collecting Digital Content, Resources, or Assets

A. Creating or Collecting New Digital Content, Resources, or Assets

A.1 Describe the digital content, resources, or assets you will create or collect, the quantities of each type, and format you will use.

A collaboration website will be created to be shared by institutions that will participate in the further development of the Computational Framework for Library and Archival Education movement. Documents including the Framework will be published on this website. Documents will include PDFs, and HTML.

A.2 List the equipment, software, and supplies that you will use to create the content, resources, or assets, or the name of the service provider that will perform the work.

We will use a hosted WordPress website at the UMD iSchool. All software used to produce content will be open source.

A.3 List all the digital file formats (e.g., XML, TIFF, MPEG) you plan to use, along with the relevant information about the appropriate quality standards (e.g., resolution, sampling rate, or pixel dimensions).

PDF, CSV, DOCx (as applicable)

B. Workflow and Asset Maintenance/Preservation

B.1 Describe your quality control plan (i.e., how you will monitor and evaluate your workflow and products).

The project team is entirely based in the U. Maryland iSchool and will convene regular meetings and follow the project timeline. Feedback and community input will be solicited and collected throughout the project and shared back with the community.

B.2 Describe your plan for preserving and maintaining digital assets during and after the award period of performance. Your plan may address storage systems, shared repositories, technical documentation, migration planning, and commitment of organizational funding for these purposes. Please note: You may charge the federal award before closeout for the costs of publication or sharing of research results if the costs are not incurred during the period of performance of the federal award (see 2 C.F.R. § 200.461).

During the project our assets will largely remain the WordPress website of collaboration. After the project period, we plan to leave these outputs on the website where they will continue to receive attention and modifications. An archival copy will also be produced and will be stored with the DCIC.

C. Metadata

C.1 Describe how you will produce any and all technical, descriptive, administrative, or preservation metadata. Specify which standards you will use for the metadata structure (e.g., MARC, Dublin Core, Encoded Archival Description, PBCore, PREMIS) and metadata content (e.g., thesauri).

N/A

C.2 Explain your strategy for preserving and maintaining metadata created or collected during and after the award period of performance.

N/A

C.3 Explain what metadata sharing and/or other strategies you will use to facilitate widespread discovery and use of the digital content, resources, or assets created during your project (e.g., an API [Application Programming Interface], contributions to a digital platform, or other ways you might enable batch queries and retrieval of metadata).

N/A

D. Access and Use

D.1 Describe how you will make the digital content, resources, or assets available to the public. Include details such as the delivery strategy (e.g., openly available online, available to specified audiences) and underlying hardware/software platforms and infrastructure (e.g., specific digital repository software or leased services, accessibility via standard web browsers, requirements for special software tools in order to use the content).

Documentation will be freely available online, through a WordPress website.

D.2 Provide the name(s) and URL(s) (Uniform Resource Locator) for any examples of previous digital content, resources, or assets your organization has created.

DCIC Center – http://dcic.umd.edu

Part III. Projects Developing Software – N/A

A. General Information

A.1 Describe the software you intend to create, including a summary of the major functions it will perform and the intended primary audience(s) it will serve.

A.2 List other existing software that wholly or partially performs the same functions, and explain how the software you intend to create is different, and justify why those differences are significant and necessary.

B. Technical Information

B.1 List the programming languages, platforms, software, or other applications you will use to create your software and explain why you chose them.

B.2 Describe how the software you intend to create will extend or interoperate with relevant existing software.

B.3 Describe any underlying additional software or system dependencies necessary to run the software you intend to create.

B.4 Describe the processes you will use for development, documentation, and for maintaining and updating documentation for users of the software.

B.5 Provide the name(s) and URL(s) for examples of any previous software your organization has created.

C. Access and Use

C.1 We expect applicants seeking federal funds for software to develop and release these products under open-source licenses to maximize access and promote reuse. What ownership rights will your organization assert over the software you intend to create, and what conditions will you impose on its access and use? Identify and explain the license under which you will release source code for the software you develop (e.g., BSD, GNU, or MIT software licenses). Explain and justify any prohibitive terms or conditions of use or access and detail how you will notify potential users about relevant terms and conditions.

C.2 Describe how you will make the software and source code available to the public and/or its intended users.

C.3 Identify where you will deposit the source code for the software you intend to develop:

Name of publicly accessible source code repository:
URL:

Part IV: Projects Creating Datasets – N/A

A.1 Identify the type of data you plan to collect or generate, and the purpose or intended use to which you expect it to be put. Describe the method(s) you will use and the approximate dates or intervals at which you will collect or generate it.

A.2 Does the proposed data collection or research activity require approval by any internal review panel or institutional review board (IRB)? If so, has the proposed research activity been approved? If not, what is your plan for securing approval?

A.3 Will you collect any personally identifiable information (PII), confidential information (e.g., trade secrets), or proprietary information? If so, detail the specific steps you will take to protect such information while you prepare the data files for public release (e.g., data anonymization, data suppression PII, or synthetic data).

A.4 If you will collect additional documentation, such as consent agreements, along with the data, describe plans for preserving the documentation and ensuring that its relationship to the collected data is maintained.

A.5 What methods will you use to collect or generate the data? Provide details about any technical requirements or dependencies that would be necessary for understanding, retrieving, displaying, or processing the dataset(s).

A.6 What documentation (e.g., data documentation, codebooks) will you capture or create along with the dataset(s)? Where will the documentation be stored and in what format(s)? How will you permanently associate and manage the documentation with the dataset(s) it describes?

A.7 What is your plan for archiving, managing, and disseminating data after the completion of the award-funded project?

A.8 Identify where you will deposit the dataset(s):

Name of repository:
URL:

A.9 When and how frequently will you review this data management plan? How will the implementation be monitored?