AFTERWORD: Towards a new Discipline of Computational Archival Science (CAS)

Richard Marciano, University of Maryland

As an afterword to this timely edited collection on Archives, Access and AI, I thought it might be helpful to amplify and extend some of the salient threads that were broached on the intersection of technology and archives.

In her introductory editorial chapter, Lise Jaillant captures three main challenges faced by cultural heritage organizations: (1) dealing with scale, (2) unlocking "dark" archives, and (3) addressing the skills gap in data science and AI. In the process, she invites contributions that not only showcase compelling interdisciplinary case studies but also summon theoretical insights.

Let me expand on these three challenges, with a concrete example from my own teaching of computational techniques to library and information science graduate students. I will make this point by highlighting a specific cultural collection: a single digitized Historical City Directory ("Post Office Directories" in the UK) for the city of Charlotte, North Carolina. City Directories are an important source of genealogical information as they were often published annually. They also supplement Census data and other local records.

1. Revisiting the Computational Challenges Faced by Cultural Heritage Organizations

a. Dealing with scale

Scale in digitized cultural heritage collections should no longer come as a surprise. The city of Charlotte (North Carolina) 1911 Historical City Directory book, as scanned and OCR-ed by the Internet Archive, yields close to 2GB of data (Charlotte itself comprises a timeseries of 62 Directories spanning an 89-year period). Directories for the entire state of North Carolina cover over 100 cities, spanning over a 100-year period (from 1860 to 1969), with close to 1,000 directories in aggregate. This represents up to 2TB (Terabytes) of digital content for the state of North Carolina alone. A rough extrapolation to the entire United States, potentially leads to 100TB of data [or two hundred 500GB hard-drives], thus merely an order of magnitude under a 1PB (Petabyte) of data. Hence, **cultural**

collections are inherently "big data." When interconnecting city directories to intersecting historical collections such as Sanborn Fire Insurance Maps, Census data, vital records, redlining data, etc., we very quickly enter the Petabyte (PB) range.

Fig 8.1: 1911 City Directory for Charlotte, North Carolina (screen snapshot from the Internet Archive).

It is in this context of scale that scalability comes into play, or the ability of archival systems to handle a growing amount of information and processing. This emphasizes how the methods that might apply to small archival holdings may not be applicable to very large holdings: "Entrat" the conversation on *Applying AI to Archives*. Applications of AI and ML to archival collections are beginning to emerge, but as Lise Jaillant highlights there is still "a lack of compelling case studies" in this space.

b. Unlocking "dark" archives

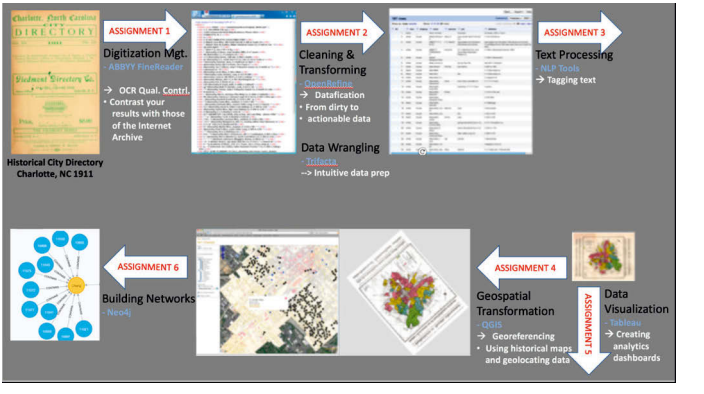
If we accept the premise of cultural collections as "big data," the next natural step is to consider the use of computational treatments to unlock "dark" cultural archives. The challenge is to see if and how we can harness the best and latest advances in data science and demonstrate its usefulness and applicability to digital cultural assets.

It is worth reflecting on the term "dark archives" which is used throughout this book. In archival literature, dark archives have a specific designation. Per the

SAA Dictionary of Archives Terminology[1] they signify either "a repository that stores archival resources for future use but is accessible only to its custodian" or "a collection of materials preserved for future use but with no current access." The US National Archives 2017 Digital Strategy[2] plan for instance, discusses a digital records infrastructure capable of safely and securely preserving several Petabytes of data in their tape-based Dark Archive, with associated descriptive metadata. There are even gradations in the literature, introducing "light archives" and "dim archives", indicating intermediate levels of access. The way "dark" archives appear to be used in this book is in the context of using AI to improve accessibility.

Back to our example, we illustrate data science driven approaches to unlocking "dark" archives by interrogating the Internet Web Archive **[relates to 2. Bell web archives paper]**, and moving the 1911 Historical City Directory pages of Charlotte, North Carolina through two processing pipelines: (1) datafication [the top part of Fig 8.2], and (2) data analysis [the bottom part of Fig 8.2].

Fig 8.2: 1911 City Directory for Charlotte, North Carolina (screen snapshot from Richard Marciano's class syllabus).



For datafication, students are asked to go inside and steer what are too often considered "black box" processes **[relates to 8. Gooding black box paper]**

1 Society of American Archivists, Definition of Dark Archives, URL: https://dictionary.archivists.org/entry/dark-archives.html [last accessed: April 5, 2021].

2 The National Archives (UK), Digital Strategy, 03.2017, URL: https://www.nationalarchives.gov.uk/documents/the-national-archives-digital-strategy-2017-19.pdf [last accessed: April 5, 2021].

including: (1) digitization (image to unstructured text, i.e. the Optical Character Recognition ABBYYFineReader tool), (2) cleaning & transforming (unstructured to structured text, i.e. the data wrangling OpenRefine and Trifacta tools), and (3) text processing (Natural Language Processing/Named Entity Recognition text tagging, i.e. GATE/ANNIE NLP/NER tool).

For data analysis, the resulting enhanced structured text is ready to be: (4) represented spatially through the creation of maps (i.e. QGIS geographical information system tool), (5) visualized interactively through the creation of analytics dashboards (i.e. Tableau data analytics tool), and (6) modeled through social networks (i.e. NoSQL Neo4j graph database). AI and ML are experienced through steps 2. and 3. We also contrast printed and handwritten text extraction approaches **[relates to 6. Hodel HTR paper and 7. Terras HTR survey paper].**

This two-phased processing pipeline is meant to provide experiential learning pathways and demonstrate the meaning of unlocking "dark archives" through the creation of an iterative "Archives, Access and Artificial Intelligence" automation workflow **[relates to 3. Jaillant design thinking paper]**

What seems equally important is to train students to think beyond "dark archives" as well and give them exposure to "dark AI" **[relates to Lise Jaillant's discussion on the "Threat of Dark AI"]**. AI cannot be examined in isolation and needs to be contextualized within the entire records management. A striking illustration, is provided by Dr. Lyneise Williams, founder of the VERA Collaborative (Visual Electronic Representations in the Archive).[3] She provides a compelling case study of how the use of photograph digitization in particular can amplify marginalization or erasure, whether through limitations in the original source documents or limitations within the technologies. The latter may unintentionally obscure visual and written features, especially those related to race, gender, and/or class. Williams offers an art historical perspective on this phenomenon, demonstrating that technical limitations can lead to erasure and distortion of archival records involving underrepresented and/or marginalized communities.[4] If marginalized people are being erased from historical records, there is not much hope for AI/ML to change these outcomes. An open challenge that arises from this work is how AI and ML

3 See https://veracollaborative.com [last accessed: April 5, 2021].

4 Lyneise Williams, What Computational Archival Science Can Learn from Art History and Material Culture Studies, 12.12.2019, in: *2019 IEEE International Conference on Big Data*, Los Angeles, CA, URL: https://ai-collaboratory.net/wp-content/uploads/2020/02/Williams.pdf [last accessed: April 5, 2021].

approaches might help uncover hidden knowledge and/or mitigate erasures within archival collections related to racial erasure.[5]

c. Addressing the skills gap in data science and AI

While this book emphasizes training humanities researchers in quantitative and computational techniques, there are two other significant dimensions I would like to highlight: (a) Establishing a framework for thinking computationally when working with digital archives, and (b) Developing interdisciplinary collaboration team building skills:

- In a recently funded IMLS Symposium grant called CT-LASER, we explored developing a Framework for Mapping Computational Thinking (CT) to Library and Archival Science Education & Research (LASER)[6], explicitly using a set of computational practices covering: (1) data, (2) modeling and simulation, (3) computational problem solving, and (4) systems thinking. CT is a form of problem solving that uses modeling, decomposition, pattern recognition, abstraction, algorithm design, and scale.[7] We provided a summary of these twenty-two CT practices spread across these four practice verticals, and we demonstrated the remapping of these concepts to archival science.[8]

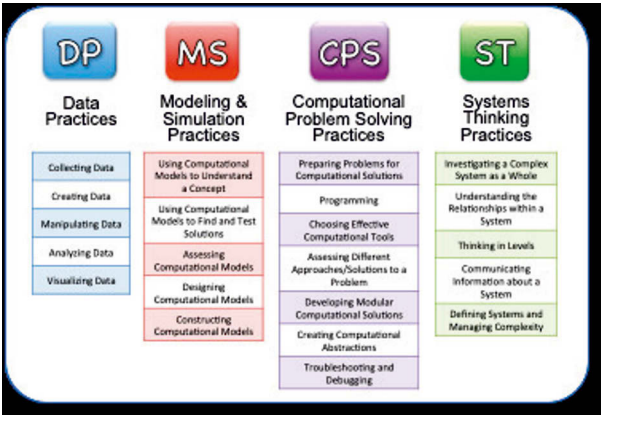
5 Lori A. Perine et al., Computational Treatments of the Legacy of Slavery (CT-LoS) "Reasserting Erased Memory," 12.12.2020, in: *2020 IEEE International Conference on Big Data*, Atlanta, in: https://ai-collaboratory.net/wp-content/uploads/2020/11/Perine.pdf [last accessed: April 5, 2021].

6 CT-LASER, final report, 01.10.2020, URL: https://ai-collaboratory.net/wp-content/uploads/2020/11/Final_Report_r.pdf [last accessed: April 5, 2021].

7 Jeannette M. Wing, Computational Thinking, in: *Communications of the ACM*, 49 (3/2006), 33–35, URL: https://www.cs.cmu.edu/ 15110-s13/Wing06-ct.pdf [last accessed: April 5, 2021].

8 Richard Marciano et al., Reframing Digital Curation Practices through a Computational Thinking Framework, 11.12.2019, in: *2019 IEEE International Conference on Big Data*, Los Angeles, CA, URL: https://ai-collaboratory.net/wp-content/uploads/2020/04/ReframingDC-UsingCT_final.pdf [last accessed: April 5, 2021].

Fig 8.3: Computational thinking taxonomy now mapped to working with digital archives (Screen snapshot from CT-LASER workshop talk, Apr. 2019).



More fundamentally, the project is addressing the integration of 'computational thinking' and 'archival thinking', as record-keeping innovation and technological development can only progress hand in hand. There is a need to accelerate opportunities for knowledge exchange and interdisciplinary synergies that will enable the infusion of archival concepts, principles, theories and methods with the computational and vice versa.[9]

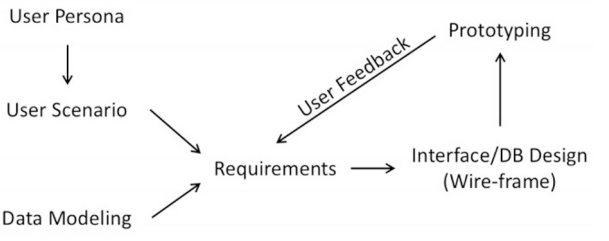
- At the University Maryland iSchool, from 2015 to 2020, I co-directed a digital curation innovation center, predicated on teaching students to work in interdisciplinary teams. During this period, we mentored over 300 students, across dozens of cultural and infrastructure projects, with a particular focus on big records and archival analytics with a mix of students with diverse backgrounds (humanities, information management, human computer interface,

9 William Underwood et al., Introducing Computational Thinking into Archival Science Education, in: *Proceedings of IEEE Big Data Conference 2018*, CAS Workshop, Seattle, WA, see 2761-2765, URL: https://ai-collaboratory.net/wp-content/uploads/2020/03/1.Underwood.pdf [last accessed: April 5, 2021].

library and archives.)[10-11] We observed that the exposure to learning how to work collaboratively in interdisciplinary teams was an indispensable skill that needed to be nurtured and developed early on.[12] The book chapter on Photoarchives **[relates to 1. Han AI applied to Photoarchives paper]** eminently illustrates the diversity of AI and Cultural Data collaborative teams, as it features five researchers with backgrounds in operations research and information engineering, mathematics, digital art history, statistics, and electrical and computer engineering.

In addition, and in support of Lise Jaillant's chapter on design thinking, our collaborative training approach emphasized an iterative design process in which ideation, prototyping, and testing are central. This relates to agile development and its iterative benefits that seem to be a natural fit with for how cultural materials are interrogated.

Fig 8.4: Iterative design process used in the Human Face of Big Data.



10 Student-Led "Datathon" Exploring Data, Investigating Methodologies, 28-29.10.2019, URL: https://ai-collaboratory.net/projects/ct-los/student-led-datathon-at-the-maryland-state-archives/ [last accessed: April 5, 2021].

11 Resistance at Tule Lake: A Conversation with the Filmmaker and iSchool Digital Curators (and Film Viewing), URL: https://ai-collaboratory.net/projects/ct-ja_ww2_camps/digital-curation-students-and-filmmaker-event/ [last accessed: April 5, 2021].

12 P. Nicholas et al., Establishing a Research Agenda for Computational Archival Science through Interdisciplinary Collaborations between Archivists and Technologists, in: *SAA 2020 Research Forum* (accepted for publication).

2. Towards a New Discipline of Computational Archival Science (CAS)

We posit the emergence of a new praxis we call *Computational Archival Science*. No one would dispute at this point in time the legitimacy of the fields of *Computational Social Science* ("Investigating social and behavioral relationships and interactions through: social simulation, modeling, network analysis, and media analysis"[13]), and *Computational Biology* ("The science of using biological data to develop algorithms or models to better understand biological systems"[14] Wikipedia). The latest addition to this computational turn may be *Computational Journalism* ("Finding and telling news stories, WITH, BY, or ABOUT algorithms"[15]).

For decades, archivists have been appraising, preserving, and providing access to digital records by using archival theories and methods developed for paper records. However, production and consumption of digital records are informed by social and industrial trends and by computer and data methods that show little or no connection to archival methods. As a matter of investigation, we have been exploring the foundations of CAS for the last five years. We captured this inquiry in a foundational paper that discusses the need to reexamine the theories and methods that dominate records practices, where we felt that this situation called for a formal articulation of a new trans-discipline, which we called *Computational Archival Science* (CAS).[16]

In this paper, our *working definition of CAS* is:

> A transdisciplinary 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.
>
> The intent is to engage and undertake research with archival materials as well as apply the collective knowledge of computer and archival science to understand the ways that new technologies change the generation, use, storage, and preservation of records and the implications of these changes for archival functions and

13 https://en.wikipedia.org/wiki/Computational_social_science [last accessed: April 5, 2021].

14 https://en.wikipedia.org/wiki/Computational_biology [last accessed: April 5, 2021].

15 Nicholas Diakopoulos, Cultivating the Landscape of Innovation in Computational Journalism, CUNY Whitepaper, 04.2012, URL: http://cdn.journalism.cuny.edu/blogs.dir/418/files/2012/04/diakopoulos_whitepaper_systematicinnovation.pdf [last accessed: April 5, 2021].

16 Richard Marciano et al., *Archival Records and Training in the Age of Big Data*, in: J. Percell et al. (eds.), *Re-Envisioning the MLS: Perspectives on the Future of Library and Information Science Education*, Somerville, MA, 2018, 179-199, URL: https://ai-collaboratory.net/wp-content/uploads/2020/10/Marciano-et-al-Archival-Records-and-Training-in-the-Age-of-Big-Data-final.pdf [last accessed: April 5, 2021].

the societal and organizational use and preservation of authentic digital records. This suggests that computational archival science is a blend of computational and archival thinking.

Archival Concepts	Computational Methods
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 in terms of why, who and how	Abstraction and ontology construction
Appraisal	File Format Characterization, File Format policies, Bulk extractor (Identifies PII), Content Preview, Tagging
Classification of archival images	AI, Line detection, image segmentation
Recordkeeping	Auto-categorization, auto-classification, e-discovery, machine learning
Personally Identifiable Information (PII)	NLP, NER, sentiment analysis
Structured data interfaces to archival materials	APIs for cultural heritage materials, graph databases
Decentralized recordkeeping	Blockchain, secure computing, trustworthiness

This approach resonates with Lise Jaillant's discussion of *"AI for Good,"* where she highlights the value of developing AI in the context of archival principles including: respect des fonds, appraisal, authenticity, and original order.

We continue to explore the mapping of archival concepts to computational methods. Papers presented at our second 2017 CAS Workshop provided evidence of the following connections:

For more information on the body of work emerging from this CAS initiative, we invite the reader to explore our CAS portal (https://ai-collaboratory.net/cas), which now features five international IEEE Big Data workshops, over 30 workshops since 2016, and over 50 research papers and presentations.

In "Computational Thinking in Archival Science Research and Education,"[17] Bill Underwood examines noteworthy archival research projects and describes how we were able to identify instances of all twenty-two CT Practices from Fig 8.3.

17 William Underwood/Richard Marciano, Computational Thinking in Archival Science Research and Education, 11.12.2019, in: *2019 IEEE International Conference on Big Data*, Los Angeles, CA, in: https://ai-collaboratory.net/wp-content/uploads/2021/03/Underwood.pdf [last accessed: April 5, 2021].

3. On the Need to Create a Network of Practitioners and Scholars in CAS

In the context of a 2019-2020 AHRC-funded International Research Collaboration Network in Computational Archival Science (IRCN-CAS) between the U. Maryland, King's College London, the Maryland State Archives, and The National Archives (UK),[18] we further observed that: (1) The new ways in which the public and researchers wish to engage with archival materials, are disrupting to traditional archival theories and practices, (2) The application of computational methods and tools to the archival problem space needs to be further explored, and (3) The contextualization of records also needs to be explored, whether through: capturing metadata, enhancing records by semantic tagging, and linking records with other records.

This led us to conclude that the way forward would benefit from establishing an international computational network for librarians and archivists.[19] This prompted us to launch the AIC Collaboratory at The Alan Turing Institute in London, UK on January 20, 2020 at the CAS Symposium held there, bringing together partners from leading academic and cultural institutions from six continents, with explicit goals to: (1) EXPLORE the opportunities and challenges of "disruptive technologies" for archives and records management (digital curation, machine learning, AI, etc.), (2) LEVERAGE the latest technologies to unlock the hidden information in massive stores of records, (3) PURSUE multidisciplinary collaborations to share relevant knowledge across domains, (4) TRAIN current and future generations of information professionals to think computationally and rapidly adapt new technologies to meet their increasingly large and complex workloads, and PROMOTE ethical information access and use.

4. On the Need to Pilot a Collaborative Network for Integrating Computational Thinking into Library and Archival Education and Practice

Finally, in response to the Editor's comment on the "lack of compelling case studies" and the need to develop real-world examples within the academic or professional literature, we conclude this afterword on a collaborative case study note, inviting

18 https://computationalarchives.net/ [last accessed: March 23, 2021].

19 Richard Marciano et al., Establishing an International Computational Network for Librarians and Archivists, in: *iConference 2019 Blue Sky Papers series*, URL: http://hdl.handle.net/2142/103139 [last accessed: April 5, 2021].

the readers of this book to consider joining forces on an initiative meant to address these gaps.[20]

To support the development of shared and reusable case studies, AIC researchers recently launched a 2-year IMLS-funded grant to pilot an online collaborative network of educators and practitioners to enable the sharing and dissemination of computational case studies and lesson plans through a Jupyter Notebook interactive computational learning platform, called *CASES* [Computational Archival Science Educational System], see: https://cases.umd.edu.

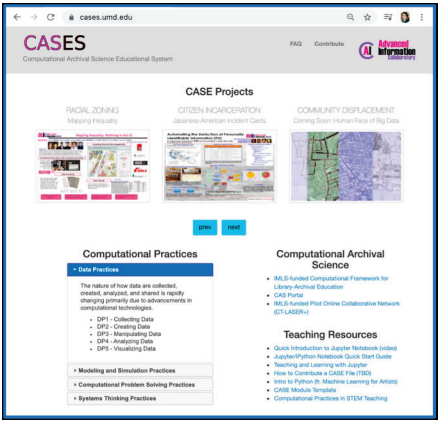
This virtual network is really a network of networks with seventeen collaborators dedicated to mapping Computational Thinking to Archival and Library practices. This Network includes: (1) a *Core Network (CN)* of seven experts in digital archives, lesson plan evaluation, project management, computational thinking, library software integration, and ethics and representation in digital collections, (2) an *Educator Network (EN)* of four educators from MLIS programs (at all ranks), and (3) a *Practitioner Network (PN)* of seven librarians / archivists representing four diverse and under-represented American collections of African-, Asian-, and Puerto Rican -American lineage: (a) the Maryland State Archives *Legacy of Slavery Project*, (b) the Spelman College *Department of Drama and Dance Photographs*, (c) Densho's *WWII Japanese American Camps Collections*, and (d) the 2019 *Puerto Rican Summer Protests ("RickyRenuncia")*. We are calling this cluster of *Practitioner Network* collections ***"Representing America,"*** to emphasize its significance and impact of training future MLIS students and exposing them to the full diversity of the American experience. In addition, we will seek feedback from an *Advisory Network (AN)* consisting of: (1) five US experts [three Practitioners at Cultural Institutions: Smithsonian National Museum of American History, Harvard Library, the US Holocaust Memorial Museum, and *two iSchool Educators* from UCLA and Drexel], and (2) International experts from all six continents. This pilot network will lead to the publication of shared, interactive, and reusable case studies which will include AI/ML exemplars, and will need to be extended and sustained through larger networks of practitioners and scholars.

To accelerate the development of case studies in Archives using AI and ML in particular, we have launched a FARM Initiative on the Future of Archives and Records Management (see: https://ai-collaboratory.net/details-aic-farm-initiative/) which seeks to develop Jupyter Notebook-based additions to the CASES repository.

---

20 Piloting an Online National Collaborative Network for Integrating Computational Thinking into Library and Archival Education and Practice, URL: https://www.imls.gov/sites/default/files/project-proposals/re-246334-ols-20-full-proposal.pdf [last accessed: April 5, 2021].

Fig 8.5: The CASES website showing a carousel of notebooks for browsing (website screen snapshot).



It is through all of these types of community intervention that we believe rapid and meaningful progress will be achieved in creating enhanced digital scholarship predicated on the integration of archives, access, and AI.

Bibliography

CT-LASER, final report, 01.10.2020, URL: https://ai-collaboratory.net/wp-content/uploads/2020/11/Final_Report_r.pdf [last accessed: April 5, 2021].

Diakopoulos, Nicholas, Cultivating the Landscape of Innovation in Computational Journalism, CUNY Whitepaper, 04.2012, URL: http://cdn.journalism.cuny.edu/blogs.dir/418/files/2012/04/diakopoulos_whitepaper_systematicinnovation.pdf [last accessed: April 5, 2021].

Lee, Myeong, et al., Heuristics for Assessing Computational Archival Science (CAS) Research: The Case of the Human Face of Big Data Project, 12.2017, in: *IEEE Big Data 2017*, Boston, MA, see 2262-2270, URL: https://ai-collaboratory.net/wp-content/uploads/2020/04/Myeong_Lee.pdf [last accessed: April 5, 2021].

Marciano, Richard, et al., *Archival Records and Training in the Age of Big Data*, in: J. Percell et al. (eds.), *Re-Envisioning the MLS: Perspectives on the Future of Library and Information Science Education*, Somerville, MA, 2018, 179-199, URL: https://ai-collaboratory.net/wp-content/uploads/2020/10/Marciano-et-al-Archival-Records-and-Training-in-the-Age-of-Big-Data-final.pdf [last accessed: April 5, 2021].

Marciano, Richard, et al., Reframing Digital Curation Practices through a Computational Thinking Framework, 11.12.2019, in: *2019 IEEE International Conference on Big Data*, Los Angeles, CA, URL: https://ai-collaboratory.net/wp-content/uploads/2020/04/ReframingDC-UsingCT_final.pdf [last accessed: April 5, 2021].

Perine, Lori A., et al., Computational Treatments of the Legacy of Slavery (CT-LoS) "Reasserting Erased Memory," 12.12.2020, in: *2020 IEEE International Conference on Big Data*, Atlanta, in: https://ai-collaboratory.net/wp-content/uploads/2020/11/Perine.pdf [last accessed: April 5, 2021].

Piloting an Online National Collaborative Network for Integrating Computational Thinking into Library and Archival Education and Practice, URL: https://www.imls.gov/sites/default/files/project-proposals/re-246334-ols-20-full-proposal.pdf [last accessed: April 5, 2021].

Resistance at Tule Lake: A Conversation with the Filmmaker and iSchool Digital Curators (and Film Viewing), URL: https://ai-collaboratory.net/projects/ct-ja_ww2_camps/digital-curation-students-and-filmmaker-event/ [last accessed: April 5, 2021].

Society of American Archivists, Definition of Dark Archives, URL: https://dictionary.archivists.org/entry/dark-archives.html [last accessed: April 5, 2021].

Student-Led "Datathon" Exploring Data, Investigating Methodologies, 28-29.10.2019, URL: https://ai-collaboratory.net/projects/ct-los/student-led-datathon-at-the-maryland-state-archives/ [last accessed: April 5, 2021].

The National Archives (UK), Digital Strategy, 03.2017, URL: https://www.nationalarchives.gov.uk/documents/the-national-archives-digital-strategy-2017-19.pdf [last accessed: April 5, 2021].

Underwood, William, et al., Introducing Computational Thinking into Archival Science Education, in: *Proceedings of IEEE Big Data Conference 2018*, CAS Workshop, Seattle, WA, see 2761-2765, URL: https://ai-collaboratory.net/wp-content/uploads/2020/03/1.Underwood.pdf [last accessed: April 5, 2021].

Williams, Lyneise, What Computational Archival Science Can Learn from Art History and Material Culture Studies, 12.12.2019, in: *2019 IEEE International Conference on Big Data*, Los Angeles, CA, URL: https://ai-collaboratory.net/wp-content/uploads/2020/02/Williams.pdf [last accessed: April 5, 2021].

Wing, Jeannette M., Computational Thinking, in: *Communications of the ACM*, 49 (3/2006), 33–35, URL: https://www.cs.cmu.edu/ 15110-s13/Wing06-ct.pdf [last accessed: April 5, 2021].