Establishing an International Computational Network for Librarians and Archivists

Richard Marciano[1], Victoria Lemieux[2], Mark Hedges[3], Yoichi Tomiura[4], S. Katuu[6], Jane Greenberg[5], William Underwood[1], Katrina Fenlon[1], Adam Kriesberg[1], Mary Kendig[1], Greg Jansen[1], Phil Piety[1], David Weintrop[1], and Michael Kurtz[1]

[1] University of Maryland, College Park MD 08544, USA
[2] University of British Columbia, Vancouver BC V6T 1Z1, Canada
[3] King's College London, London WC2R 2LS, UK
[4] Kyushu University, Fukuoka 819-0395, Japan
[5] Drexel University, Philadelphia PA 19104, USA
[6] University of South Africa, RSA

marciano@umd.edu, v.lemieux@ubc.ca, mark.hedges@kcl.ac.uk, tom@inf.kyushu-u.ac.jp, skatuu@gmail.com, jg3243@drexel.edu, underwod@umd.edu, kfenlon@umd.edu, akriesbe@umd.edu, mkendig@umd.edu, jansen@umd.edu, ppiety@umd.edu, weintrop@umd.edu, mkclandcats@verizon.net

Abstract. Research and experimentation are underway in libraries, archives, and research institutions on various digital strategies, including computational methods and tools, to manage "Collections as Data" [1]. This involves new ways for librarians and archivists to manage, preserve, and provide access to their digital collections. A major component in this ongoing process is the education and training needed by information professionals to function effectively in the 21st century.

Accessible and transferable infrastructure is a key requirement in creating a network of collaboration for information professionals to fully realize the full potential of managing "Collections as Data." Elements needed include:

1. Open source research and educational platforms to remove barriers to access to curation tools and resources. These are needed to deliver and share computational educational programs.
2. Creation of a Cloud-based student-learning environment.
3. Development of Open Source software architectures that use computational infrastructure.
4. Exploration of new pedagogies for educating librarians and archivists in computational methods and tools.
5. Establishment of a community of practice for developing collaborative projects, and liaising with the wider international iSchool community and practitioners in the field.

Our "Blue Sky" proposal seeks to explore a number of these challenges (infrastructure, computation, collaboration, learning) that stimulate the iSchool research community and have the potential to jumpstart international collaborative networks.

A significant outcome would be the development of an international computational network for supporting librarians and archivists, akin to the existing Sloan Foundation funded "Data Curation Network", which over the next five years seeks to model a cross-institutional staffing approach for curating research data in digital repositories [2].

Keywords: Computational Archival Science, Collections as Data, Computational Network for Librarians and Archivists, Digital Curation, Archival Analytics.

1 Background

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 [3]."*

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 [4].

The U. Maryland iSchool is an active member of this community and has furthered the research and curricula development with computational treatments of collections [5]. This can be seen through their work on Computational Archival Science (CAS) [6], defined as:

> *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, 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 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," [7] 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." [8]

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. [9]

The Smithsonian Institution has conducted cutting-edge work showing how computational treatments using artificial intelligence can revolutionize archival museum research [10]. 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? (2) With what accuracy can machine learning algorithms recognize members of two similar plant families? Preliminary results are outstanding.

Finally, The National Archives (TNA) in the UK has recently hosted two workshops on automating the archive. At the September 7, 2018 workshop held at King's College London, the focus of the workshop was on Computational Archival Science and Automating the Archive [11]. This event explored how computational approaches can be used to support archival practice in the creation and preservation of reliable and authentic records and archives, taking into account users of archives, and how access and interaction can be supported and enhanced. At the September 4, 2018 forum on Artificial Intelligence and Archives, the focus was on the use of AI at the service of archival appraisal, selection and sensitivity review of born digital records might, and the emergence of machine learning technologies as radical new capabilities and new possibilities for archival processing [12].

2 Integrating iSchool Research

The authors of this "Blue Sky" paper have explored this topic for the last two and a half years and have created an initial international network with iSchool researchers from the UK, Canada, the US, and Japan. The goal is to build on research topics that are unique to iSchools and share advances, challenges, and research ideas and oppor-

tunities at the larger iConference, broadening the current agenda, and also exploring funding and sustainability approaches.

We view this "Blue Sky" approach as the kind of visionary and integrative agenda needed to help integrate research within iSchools (potentially bringing together more traditional Libraries and Archives research ideas with emerging Computer Science, HCI, and Information Science advances), and also across iSchools with varying foci and audiences.

References

1. Padilla, T.:Mellon Foundation Awards $750K Grant To Support Collections as Data (2016), https://www.library.unlv.edu/newsblog/2018/07/mellon-foundation-awards-750k-grant-support-collections-data.html
2. Data Curation Network: https://sites.google.com/site/datacurationnetwork/
3. Padilla, T, et al. (2016): *"Always Already Computational: Library Collections as Data"*. Link: https://www.imls.gov/grants/awarded/lg-73-16-0096-16
4. ACM Journal on Computing and Cultural Heritage (JOCCH): https://jocch.acm.org/index.cfm
5. The "Computational Archival Science (CAS) Portal". Link: http://dcicblog.umd.edu/cas/
6. 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.
7. Collections as Data: Stewardship and Use Models to Enhance Access, 9/27/16). http://digitalpreservation.gov/meetings/dcs16.html
8. Collections as Data: IMPACT (Jul. 25, 2017). http://digitalpreservation.gov/meetings/asdata/impact.html
9. "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
10. Smithsonian Magazine, Nov. 3, 2017. Link: https://www.smithsonianmag.com/smithsonian-institution/how-artificial-intelligence-could-revolutionize-museum-research-180967065/
11. TNA and KCL Workshop on Computational Archival Science: Automating the Archive, Sep. 7, 2018, The National Archives, Kew, UK. See: http://dcicblog.umd.edu/cas/9-7-2018_uk-nara-kcl_cas-workshop/
12. International Council on Archives (ICA) Forum of National Archivists: AI and Archives. Sep. 4, 2018, The National Archives UK, Kew. See: https://www.ica.org/sites/default/files/ai_and_archives_-_digital_symposium_flyer_-_pdf.pdf