Marciano, R., Lemieux, V., Hedges, M., Esteva, M., Underwood, W., Kurtz, M. & Conrad, M. (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.

ARCHIVAL RECORDS AND TRAINING IN THE AGE OF BIG DATA

Richard Marciano[1], Victoria Lemieux[2], Mark Hedges[3], Maria Esteva[4], William Underwood[1], Michael Kurtz[1] and Mark Conrad[5]

[1] U. of Maryland iSchool,
[2] U. British Columbia iSchool,
[3] King's College London, Dep. Of Digital Humanities,
[4] Texas Advanced Computing Center,
[5] National Archives and Records Administration

ABSTRACT

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. The purpose of this chapter is to reexamine the theories and methods that dominate records practices. The authors believe that this situation calls for a formal articulation of a new transdiscipline, which they call computational archival science (CAS).

Keywords: Computational archival science; CAS; archival thinking; computational thinking

INTRODUCTION

For decades, archivists have been appraising, preserving, and providing access to digital records 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.

If archivists are to successfully adapt to the current environment, they must examine the theories and methods that dominate records practices. At the same time, researchers are struggling with issues of archiving and sharing data that can be addressed by reference to archival theories and methods. While there are incipient signs of change in the archival space, we believe that this situation calls for a formal articulation of a new transdiscipline called computational archival science (CAS), which we suggest may be defined as:

Proposed Working Definitions of Computational Archival Science (CAS):

> *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 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.*

The aforementioned definition is, by nature of its novelty, still tentative and evolving. For example, we recognize that in a transdiscipline there will be a two-way exchange of knowledge between the foundational disciplines that the current definition does not adequately reflect.

After making a case for this new transdiscipline, we present case studies that demonstrate the changing environment of archival practices and some examples of interdisciplinary efforts to address these changes. The case studies demonstrate how current theories and methods of one or more of the disciplines could benefit from those of one of the others, successful collaborations between disciplines, and early practitioners of CAS. At the end of each case study, we discuss possible topics, takeaways, and methods of incorporating CAS into MLS education to better address the needs of today's MLS graduates looking to employ "traditional" archival principles in conjunction with computational methods.

THE CASE FOR A NEW TRANSDISCIPLINE

It is relatively easy to make the case for the application of computational methods and resources to large-scale records and archives processing, analysis and storage, long-term preservation, and access. Case in point, the Networking and Information Technology Research and Development (NITRD) Program coordinates federal research and development investments in advanced digital technologies. The NITRD Program's Supplement to the President's FY 2017 Budget (NITRD 2016) emphasizes topics of data preservation; analysis of large, heterogeneous collections and records; scalable ingest; machine reading of records; trustworthiness and reliability of automated systems; document summarization/distillation; automatic content extraction; effective analytical tools for decision makers; data capture, curation, management, and access; and data privacy, security, and ethics. In addition, the May 2016 White House Federal Big Data Research and Development Strategic Plan (Marzullo, 2016) discusses the creation of next-generation capabilities; the understanding of trustworthiness of data and resulting knowledge; the increase in the value of data through policies; the understanding of big data collections with regard to privacy, security, and ethics; and the creation of big data benchmarking centers.

In an era of abundant digital data, archivists are increasingly hard-pressed to manage archives without such techniques. Moreover, these techniques offer many advantages that traditional archival approaches do not. Take, for example, the potential of data sciences such as information visualization to transform archival preservation processes (Weijia, Esteva, Jain, & Jain, 2011) or the opportunity to apply visual analytics to transform archival representations and finding aids from two-dimensional hierarchies into multidimensional graphical representations complete with interactivity that supports integration of human and machine analysis. Examples of the application of information visualization and visual analytics techniques to archival description can be found in Lemieux's (2012) survey article on the topic, including work by Robert Allen, Ian Anderson, and Mitchell Whitelaw. Their research makes contributions both to archival science and to information visualization and visual analytics, itself a multidisciplinary field. Lemieux's (2014) work on "third order" archival systems not only posits the use of novel visual representations and visual analytics as the basis of new intelligent finding aids – as well as drawing insights from her visual analytics research on the nature of the record and archival provenance – but also makes a contribution to visual analytics by advancing knowledge and application of visual analytics to ontology visualization. There is even a case to be made for, and evidence of, the application of computational theories and methods to expand archival theories and methods. Duranti and Michetti (2015) wrote that, "The 'autonomous' internal core of archival knowledge coming to us from ancient times is very small" (p. 81). They note influences from law, philology, history, social science, library science, and information science as among those disciplines that have influenced archival thinking in the past, suggesting the need for an interdisciplinary thinking. While none of these examples requires a new combined discipline or even transdiscipline, computational methods and tools can be applied to perform archival functions without the need for any theoretical or methodological integration or synthesis in much the same way as computational methods are applied in finance without integration with financial theories and methods.

Beyond the application of technology to improve efficiency, productivity, and precision in support of traditional archival functions, or even reconsideration of archival theory in light of computational theories and methods, there is a need to fundamentally transform both disciplines in order to infuse archival theories, principles, and methods with the computational and, equally, to infuse the computational with archival conceptualizations and theories of the record. With respect to the latter, the purpose is to introduce computational science to archival science and the archival method, which focuses its attention on the record and aggregates of the record – as opposed to focusing on information or data processing as in computer science. Why would this be necessary or, indeed, valuable? Computer science is now concerning itself with areas of investigation that have long been the focus of archival science. There is no more obvious an example of this than recent research on provenance. At a multidisciplinary meeting on provenance held in 2015 (Lemieux, 2016a), researchers from the field of computer science found the archival perspective on multiple provenances as articulated by Australian archivists (Cunningham, 2016) novel and innovative. Equally, archival science researchers benefited from work within computer science articulating workflow provenance (Ludäscher, 2016). The key point here is that each field benefited from the exchange, and, through the intermixing of these perspectives, the common theory of provenance was enriched and expanded. That may be sufficient, and there may be no need to consider a combined field or discipline, save for the fact that both fields are looking to use provenance as a means of capturing and assessing the trustworthiness of information as evidence when created, processed, and stored across multiple distributed systems. Thus, they share a common core problem, which, by combining their disciplinary knowledge into a single field of vision, as it were, offers the combined disciplines the ability to go far beyond what each could achieve working on its own.

As more records are born digital, understanding the methods and context of creation and how they shape the characteristics of records cannot be achieved with reference to archival science theories and methods alone. Achieving this understanding will benefit from drawing upon computational science as the science underpinning the formation, processing, and storage of digital records. For the past 40 or so years, archival researchers have considered the impact of digital technologies on records and recordkeeping (Ambacher, 2003). With the advance of digital records and the growing complexity of information systems, the ability to shape the formation, processing, and preservation of digital records to serve the aims of forming, preserving, and making accessible trustworthy records is likely to depend on applying a deeper and more integrated blend of archival theory, principles, and methods with computational theory, principles, and methods. Neither field can serve this purpose on its own: Archival science without computer science lacks knowledge necessary to understand and effectively administer digital records; at the same time, computer science without archival science lacks knowledge needed to generate enduring and trustworthy memory and evidence – hence, the need for a combination of the two disciplines into the new transdiscipline, CAS.

To argue for CAS, suggests that neither traditional archival science should cease to exist, nor computer science – or data science, for that matter – should merge with archival science. Each discipline can, and must, continue to stand as a separate field, as each has concerns and areas of investigation that do not intersect with that of the other. After all, there will continue to be non-digital records that will require management and preservation. Equally, many computer systems will be concerned solely with information processing, with no need to design for the creation and preservation of records that can serve as trustworthy evidence.

MOTIVATING CASE STUDIES

The following eight case studies offer examples of interdisciplinary efforts to address the changing context of recordkeeping and character of records: (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. Each of the case studies concludes with a "Takeaways for CAS/MLS Education" statement.

1. Evolutionary Prototyping and Computational Linguistics

The Presidential Electronic Records PilOt System (PERPOS) was a research project sponsored by the Electronic Records Archives Program of the National Archives and Records Administration (NARA) and led by the Georgia Tech Research Institute (Underwood, 1999; Underwood, Kindl, Underwood, & Laib, 2001; Underwood et al., 2006). The primary objective of the research was to provide advanced computational technologies to support archivists in processing the personal computer records created and used in the Executive Office of the

President during the administration of President George H. W. Bush (1989–1993). The PERPOS Project was not a software development project, but a research project that used evolutionary prototyping to lead to a better understanding of the requirements for archival processing of presidential digital records by applying existing computer technologies and creating new technologies to meet these requirements. Evolutionary prototyping involves creating a software prototype supporting or performing a particular task. The users interact with the prototype and evaluate the user interface, the functionality, and the performance of the prototype. The prototype is then modified to satisfy the resulting better-understood needs and iteratively refined (Software Prototyping, 2016). This use of evolutionary prototyping is an example of using a computer science method to bridge the knowledge bases of two separate disciplines to facilitate transdisciplinary work.

This project combined computational linguistics concepts of grammars, parsing, and information extraction with the archival science concepts of documentary form and archival description. Computational linguistics is an interdisciplinary field in which computational methods are applied to linguists' theories about natural languages, the result being statistical or rule-based computational models of natural language. Archival description is "the process of analyzing, organizing, and recording details about the formal elements of a record or collection of records, such as creator, title, dates, extent, and contents, to facilitate the work's identification, management, and understanding" (Pearce-Moses, 2005, p. 25). Descriptions can be at the collection, office, series, folder, or item level.

It was found that series descriptions for paper records in the Bush Presidential Library included a description of the document types, topics, actions, recipients, and inclusive dates of the items in a series. Metadata extraction is a critical aspect of ingestion of collections into digital archives and libraries. A computational linguistics method for automatically recognizing document types and extracting metadata from digital records was developed (Underwood, 2010) based on a method for automatically annotating semantic categories such as names, job titles, dates, and postal addresses that may occur in a record. It extends this method by using the semantic annotations to identify the intellectual elements of a document's form, parsing these elements by using context-free grammars that define documentary forms and interpreting the elements of the form of the document to identify metadata such as the chronological date, author(s), addressee(s), and topic. Context-free grammars were developed for 14 of the documentary forms occurring in presidential records. In an experiment, the document-type recognizer successfully recognized the documentary form and extracted the metadata of two-thirds of the records in a series of presidential e-records containing 21 document types (Underwood, 1999).

Every U.S. federal agency is required by statute and regulation to manage its records according to a record retention schedule that requires the classification of email as records. The increasing volume of intra-office, incoming, and outgoing email makes it impractical for individuals and organizations to manually categorize and manage email. An experiment was conducted in which support vector machine classifiers were trained to classify six categories of actual Georgia Tech emails whose retention was determined by the University System of Georgia Records Retention Schedule. The six classifiers were then evaluated in classifying email not in the training set. The result was 99% accuracy in classifying 198 additional emails (Underwood & Laib, 2012). It was concluded that machine-learning techniques can be used to achieve a high degree of accuracy when automatically categorizing email that is then retained according to an organizational record retention schedule. Additional research is needed to determine whether the efficiency and level of accuracy can be maintained when the problem is scaled to hundreds of retention categories and millions of emails.

Takeaways for CAS/MLS Education

Evolutionary prototyping can be used as a collaborative research method to bring computational thinking and archival science thinking together. These case studies also illustrate how the two disciplines can come together to generate new insights and approaches. In relation to educational program design, the project points to some of the new knowledge that those working in the "archival problem space" may require, that is, knowledge of grammars, parsing, information extraction, machine learning, and topic modeling.

2. Graph Analytics, Digital Humanities and Archival Representation

Graph Analytics in Digital Humanities

Graph theory, and its manifestation in so-called non SQL (NoSQL) graph databases, is emerging as a powerful model for representing and querying complex, interconnected, and large cultural data sets. In April 2016, when the Panama Papers were revealed by a consortium of investigative journalists, from the millions of unstructured documents and terabytes of data, graph databases were used to link individuals to banks and financial institutions as well as transactions, with interactive visual searches to relate people to offshore accounts (Panama Papers, 2016). Similarly, researchers at King's College London (Blanke, Bryant, & Hedges, 2013) from the European Holocaust Research Infrastructure portal have developed a novel archival and administrative metadata framework centered around graph databases. Dubbed "Graph Archives," this tool offers a new way of linking dispersed cultural resources involving people, places, time, and events into very large social-networking-like graphs. This approach mirrors Facebook's Social Graph, Google's Knowledge Graph, and Twitter's Interest Graph. Modeling the displacement of people with graph databases allows users of the tool to answer historical questions in novel ways: who was connected to whom at a moment in time and place and what relationships and relocation patterns were present, to name just two examples.

Graph Analytics in Archives

Working independently of digital humanities scholars, archival science scholars have seen the value of graphical abstractions and representations of archival provenance. In her paper on "third order" archival systems, Lemieux (2014) describes building a prototype graphical representation of the Canadian context of financial electronic records. More recently, Thibodeau (2016) argued that graph theory can enhance the treatment of archival provenance. Graph theory, he argues, offers suitable, quantitative methods of analysis, and opens possibilities for the use of automated analytical and visualization techniques that are well suited to the objectives of records management.

Takeaways for CAS/MLS Education

As graph analytics are being used increasingly by archives users and, at the same time, are opening up new possibilities for archival representation and redefining in more precise terms basic archival and records management principles such as provenance, the teaching of archival science must adapt to reflect these changes. CAS students should learn about graph theory, graph databases, and graph analytics and how these underlying theories and technologies can be applied to represent the archives and their context of creation and as means of supporting new approaches to archival research.

3. Computational Finding Aids

The exploration of computational finding aids is an endeavor that illustrates the potential application of computational methods to large-scale archives processing in support of traditional approaches to appraisal, arrangement, and description. As per the Society of American Archivists' Glossary of Archival and Records Terminology , a finding aid is a "tool that facilitates discovery of information within a collection of records" (Pearce-Moses, 2005, p. 168). Traditionally, these approaches have largely been conducted manually.

When considering billion-record digital archives, there is no question that the process of analyzing large collections and extracting metadata will require computational methodologies and interactive visual analytics. We illustrate this changing landscape with a 100 million file collection from NARA reconstituted at the University of Maryland's Digital Curation and Innovation Center (DCIC) on an open-source archive that potentially scales to billions of records. The archive supports in situ file analysis through cyberinfrastructure that provides a library of extractor services and the ability to plug in custom extraction services (Jansen, Marciano, Padhy, & McHenry, 2016). Existing extractor services convert legacy formats, identify file formats, and extract a myriad of features (from handwritten text, typed text, historical images, face recognition, Optical Character Recognition (OCR), etc.). Extracted features are tagged back to the original record, thus becoming enriched metadata, and indexed through a scalable engine such as Elasticsearch or Solr. The resulting indexed information is conducive to the creation of faceted interfaces with the writing of queries visually and interactively, and with the creation of dynamic dashboards for different audiences.

While these resulting finding aids appear to bear little to no resemblance to traditional finding aids, one can easily use traditional archival arrangement as a search facet, allowing users to browse the hierarchical archival tree and displaying derived knowledge by level of description abstraction: from record group, to series, to file unit, to items (to reflect NARA parlance), drilling down or zooming in and out of collections. This has the potential to respect the arrangement of the fonds while revealing connections across the archival "boxes," or silos, unlike a traditional Google search. This also has the potential to transform archival content into knowledge at scale while providing enhanced analytical tools to the digital archivist of the future (Heard & Marciano, 2011; Weijia et al., 2011).

Takeaways for CAS/MLS Education

This chapter suggests that we also need to teach students about big data infrastructure – e.g., the architecture needed to support, ingest, and manage billion-file repositories such as the DCIC's – about the tools that can be used now; and how to learn about new tools, as they are constantly emerging. Instruction is required in both the theory and practical techniques of applying information visualization and visual analytics techniques to archival work and, equally important, to building interactive visual interfaces with underlying analytic capabilities such as content recognition, natural language processing, and data summarization. With the introduction of computational findings aids will come a need for educating students in the theories, principles, and techniques of human–computer interaction, as recognized as long ago as 2010 by Steve Bailey and Jay Vidyarthi, to ensure effective design of what may evolve into highly complex intelligent cognitive systems (Bailey & Vidyarthi, 2010).

4. Digital Curation

In MLS programs, courses in archival methods cover records appraisal, arrangement, description, preservation, and access. These methods are typically applied to records of government, academic, and business activities and not to records of scientific and humanities activities, which are primarily published in journals, conference proceedings, and books and are collected and managed in libraries. Until recently, it was uncommon for government, academic, and business archives to preserve and provide access to scientific and humanities research data, but these trends are now changing.

The National Science Foundation (NSF) expects investigators to share with other researchers the primary data, samples, physical collections, and other supporting materials created or gathered in the course of work performed under NSF grants. The NSF requires that proposals include a supplementary document of no more than two pages labeled "Data Management Plan." This plan may include, but is not limited to, data, publications, samples, physical collections, software, and models. Here, data management involves all stages of the digital data life cycle, including capture, analysis, sharing, and preservation. The focus is on sharing and preservation of digital research data. The data access plan must address the institutional strategy for providing access to relevant data and supporting materials. The National Institutes of Health and the Department of Energy, Office of Science, have similar requirements (Northwestern Libraries, 2017).

NASA's Planetary Data System (PDS) at the Jet Propulsion Laboratory is an example of a science data archive (Underwood, 2005). The methods of this system included data management plans for NASA's planetary science projects, peer review of scientific data sets, self-describing data formats, collections of scientific data sets, documentation and publications, software for processing data sets, and catalog descriptions of data sets. The PDS integrates computer technology with archival methods.

Neil Beagrie defined digital curation in his description of the Digital Curation Centre (Beagrie, 2006):

> The term digital curation is increasingly being used for the actions needed to maintain digital research data and other digital materials over their entire life-cycle and over time for current and future generations of users…Implicit in this definition are the processes of digital archiving and digital preservation, but it also includes all the processes needed for good data creation and management, and the capacity to add value to data to generate new sources of information and knowledge. (p. 4)

Digital curation (in some circles) is seen to extend archiving and preservation by adding value to digital objects through indexing; adding metadata, annotation, or markup of various forms, including semantic markup/ontologies (using both manual and automated methods); enhanced discovery and access (including retrieval and visualization); and facilitating interoperability and integration. However, in other circles, these extended activities are viewed as all being part of existing traditional archival methods and part of an ongoing debate. Such enhancements may support additional scientific experiments, interpretation of text by scholars, development of exhibits, and other scholarly pursuits. Digital curation is concerned with curating digital objects and information in all their varied guises. Such digital assets have a key role to play, not only in scholarly research or in the domains traditionally associated with the management of information, such as libraries, archives, and other memory institutions, but also in a much broader range of institutions and activities. Digital information is ubiquitous across organizations of all types and is a key asset for government and for industries as wide ranging as banking, law, and medicine. In particular, there are close links with the parallel fields of digital asset management, which arose in the business world within the media, and content industries to address the need to manage complex information assets within an enterprise context and to exploit their social, cultural, and commercial values.

Takeaways for CAS/MLS Education

Government sponsors of scientific research, scientific communities, and universities have begun to realize the value of scientific data as records that should be preserved and shared. To serve these communities, MLS students need to have training and experience in building and managing scientific data archives.

The Digital Curator Vocational Education Europe (DigCurV) project was an initiative funded by the European Commission to investigate and identify the requirements for vocational education in digital curation and to establish a corresponding curriculum framework. The framework reflects the understanding of digital curation as a multi-skilled profession that involves a wide range of competences (Molloy, Gow, & Konstantelos, 2014). MLS programs need to address these various aspects of digital curation and digital asset management if they are going to prepare students to take advantage of the range of opportunities open to them outside traditional library and information organizations.

5. Public Engagement with (Archival) Content

Crowdsourcing in the context of academic or memory institutions may be defined as the process of leveraging public participation in or contribution to projects and activities (Dunn & Hedges, 2013). To make use of the transformations that the web has brought to processes of collaboration and communication, a range of initiatives have been undertaken in these sectors involving public participation with a view to enhancing, augmenting, or opening up cultural material, blurring the boundaries between the spaces occupied by professional and non-professional communities and transforming the relationship between cultural organizations and the wider community.

Public involvement in archives can take many forms, ranging from enhancing digitized documents through, for example, transcription of handwritten text (Brohan et al., 2009) or geo-referencing historical maps (Fleet, Kowal, & Pridal, 2012); to activities more closely aligned with "traditional" archive or library practice, such as cataloging, or, more informally, tagging and categorizing, documents to facilitate discovery and preservation (Burgess, 2016); to more complex activities such as commenting on or discussing content, adding contextual information such as personal experiences or memories, or constructing alternative narratives and interpretations.

The activities thus range from independent microtasks that are farmed out to individual participants – a model more akin to business crowdsourcing – to participatory creation of complex information objects, or the "social curation" of information (Zarro & Hall, 2012). This is perhaps closer to the traditional meaning of "curation" within a museum or gallery context and involves the collection, organization, and dissemination of information on a particular topic of interest, as exemplified on sites such as Storify (https://storify.com) and Pinterest (https://www.pinterest.com). The Citizen Archivist Initiative at NARA is also a good example that includes transcription, tagging, uploading personal images, and collaborative editing of articles (http://www.archives.gov/citizen-archivist/).

Takeaways for CAS/MLS Education

There are many opportunities for training in a re-envisioned CAS MLS setting. Crowdsourced description has come up in the context of the imProvenance Group discussions (Lemieux, 2016a), with the key takeaway being that computational approaches to tracing the provenance of both the objects or records being curated and the mixed archivists and crowd description of them are a big data challenge requiring computational approaches to be effective and sustainable. Increasingly, hybrid approaches to big data are emerging where humans are made part of the big data challenge by both automating and weaving humans into automated processes (Marciano, 2015).

6. Authenticity

The rapid obsolescence of computing technologies creates difficulties for those concerned with the long-term preservation of records in digital form. The potential need to migrate these records across hardware and software technologies raises questions related to the records' authenticity. How can one ensure that sets of digital records have not been intentionally or inadvertently modified? How can one ensure that long-term preservation methods do not compromise the authenticity of digital records?

A formal method has been described for analyzing records management, and archival procedures and systems to determine whether they maintain and preserve authentic records over time. The analysis procedure is based on a formalization of archival and diplomatic concepts and principles as definitions and axioms. Concepts such as digital record, record series, and archival integrity are defined, and axioms characterizing authentic documents and authentic records are formulated. A procedure is described for storing and retrieving the digital records of a record creator that incorporates elements to ensure the integrity and authenticity of the records. The theories of record integrity and authenticity are used with computer science theories of communications security and belief to prove that the procedure achieves its goal of preserving the integrity and authenticity of the digital records (Underwood, 2002).

A blockchain is a distributed database that maintains a growing list of records (blocks) that are secured from unauthorized revision or tampering. Each block has a link to a previous block. The blockchain is the central element of the digital currency bitcoin, and it serves as a ledger for transactions. Blockchain technology is said to establish the authenticity of the records. This technology is also generating new "on chain" records, such as smart contracts. Computer science has given birth to this innovative new technology. But can these new blockchain records really be trusted? Ensuring the ability to ascertain, check, and audit trustworthy records is essential in evaluating blockchain technology, especially because its potential is perceived as disrupting a range of industries, including data and identity management, healthcare, insurance, and peer-to-peer economies. Some even propose that traditional archives be replaced with new blockchain-based, decentralized, autonomous archives.

With the rise of blockchain technology for recordkeeping, there is a need to develop the criteria against which the trustworthiness of the blockchain can be evaluated. Archival science theories, principles and practices, and the international standards that have been derived from them can point the way. Archival theory addresses key concepts needed to produce and preserve trustworthy records: accuracy, reliability, and authenticity intertwined with the concept of provenance.

Without archival knowledge, successful development and implementation of what is essentially a recordkeeping technology is, arguably, unlikely to achieve the much-hyped innovations and disruptions its proponents foresee and may even lead to unintended negative consequences for society, such as loss of the critical documents created and potentially stored on a chain (e.g., smart contracts) or an inability to establish the authenticity of documents claiming rights and entitlements, such as copyright and ownership of land (Lemieux, 2016a; Lemieux, 2016b; Lemieux, 2016c).

Takeaways for CAS/MLS Education

Work on the long-term preservation of authentic digital records demonstrates the continuing value of archival concepts and principles in the face of rapidly evolving records creation and management technologies. CAS must therefore remain firmly rooted in these traditions, extending and applying them to new record making and keeping contexts. At the same time, new forms of records and new contexts will iteratively feed back into the understanding of core archival theories and concepts, combining them with computational theories and concepts to derive a fundamentally integrated conceptual and theoretical foundation for CAS.

Students must be sufficiently well-versed in the technical features of new records creation and recordkeeping infrastructures, such as digital signatures, data integrity, hash codes, and blockchain, that they have the technical knowledge needed to conduct assessments of the viability of long-term preservation of authentic digital records and the IT security and other risks that may impede this goal. Students of CAS will also, therefore, need to receive training in software risk assurance in order to have the knowledge and competencies necessary to protect the integrity of the archive.

7. Confluences between Archival Theory and Computational Methods: Cyberinfrastructure and the Records Continuum

A fundamental skill for future archival professionals is to be able to navigate with ease the confluence between archival theory and the trends that dominate how digital records are used, socialized, and preserved in the present. A way of addressing such confluence is to re-interpret, hypothesize, test, map, and criticize archival theories in the light of the possibilities afforded by existing technologies and the requirements imposed by emergent record uses.

A demonstration of such confluence is the interplay between the records continuum theory and the possibilities provided by cyberinfrastructure ecosystems. One of the most provocative archival theories, the records continuum (Upward, 2005), states that the boundaries between different record functions are seamless and can be assumed at any point from the moment of records creation, as there are no fixed time-based stages for managing active records and archives. At any given point, a record that is restricted to some users may not be restricted to others or may have informational value to some and evidential to others. What changes are the contexts in which those functions exist, pointing to the diversity of uses and users of records, each with their unique needs and interests in one or more records roles. Using cyberinfrastructure as a backdrop, we can observe and analyze how the records continuum theory can be hypothesized and tested.

Cyberinfrastructure is the combination of computational facilities, software, services, and human resources within which systems of different configurations can be built to address diverse functions. Using cyberinfrastructure, (Esteva, Sweat, McLay, Xu, & Kulasekaran, 2016) designed and implemented an automated recordkeeping system that seamlessly gathers, curates, archives, and publishes data generated in an open science supercomputer. The gathered data contain information about users and about the scientific libraries they used to run jobs on a supercomputer. Throughout the automated curation process, on the one hand, the data are anonymized, integrated with other metadata, authenticated, archived, packaged with corresponding metadata, and deposited in a publicly accessible repository system. On the other hand, non-anonymized data are used to manage the supercomputer accounts and its performance.

These transformations are possible as data transits across analyses, curation and storage platforms, and services in the cyberinfrastructure. Analyzed through the prism of the records continuum, the system shows that records can have different roles and values at the same time and that each of these functions is instantiated using different and complementary computational resources – consisting of repository systems virtual machines, and databases – joined through policies, scripting, efficient networks, metadata protocols, and links that bring to different users the function of the record that they need when they need it. It also shows that the records continuum theory stands up to being tested against computational modes of records management and archiving and has very concrete and useful practical implications.

Takeaways for CAS/MLS Education

The questions for educators are: What are the knowledge, skills, and capabilities that archivists of the future need to have to create such systems, and what is the role of CAS in building those skills? The challenge is not only what to teach but, more importantly, how to teach so that students think in flexible and creative ways about theory and novel practices. CAS can contribute to these future professionals by training students with the technical and management skills required to create services within cyberinfrastructure and embed in those skills the creativity and vision encountered in archival science. In turn, the capacity to observe the reality of how people and institutions use and create information with critical and accepting eyes in context with new skills and technological knowledge will allow for re-envisioning archival and computational theory and creating services that advance and promote both fields.

8. Spatial and Temporal Analytics

Electronic records are (often) endowed with spatial and temporal characteristics that allow for unique types of analyses and linking. Revealing these attributes in archival records to provide new modes of access and understanding is becoming commonplace. In Heard and Marciano (2011), spatial attributes are extracted and computed to create geospatial indexes and graphical user interfaces to relate tens of millions of government records. In Winling, Connolly, Nelson, and Marciano (2017) and Travis et al. (2016), historical archival content is unlocked and integrated through geocoding, geo-referencing, and spatial mapping. In Dingwall, Marciano, Moore, and McLellan (2007), the temporal dimension is added by demonstrating that large quantities of geospatial data can be captured within a recordkeeping system and versioned over time for long-term preservation purposes, allowing for the development of temporal cloud interfaces to spatial content. Indeed, funding agencies such as the National Historical Publications and Records Commission (NHPRC) have funded investigations in the preservation of geospatial e-records: the e-Legacy project in 2007 with R. Marciano, principal investigator (investigating the appraisal, accessioning, and preservation of the geospatial e-records of the state of California), and the ICAP project in 2003 with R. Marciano, principal investigator (investigating the ability to compare versions of records and run historical queries). In addition, 3D printing is emerging as a technique to represent and interact with spatial content (Clark, McKeon, Marciano, & Bailey, 1998).

Takeaways for CAS/MLS Education

This case study suggests that there is great potential in teaching students about the management of temporal and geospatial data. Exposure to processing of events and spatial relationships creates a foundation for digital engagement and storytelling and allows for new modes of computing and analyzing archival content.

TEACHING OPPORTUNITIES FOR THE NEW MLS

Lessons learned from these case studies are likely to elicit CAS building block topics to create innovative classes that emphasize new modes of collaboration and interdisciplinary work. We look at blending elements of archival thinking and computational thinking, a form of problem solving that uses modeling, decomposition, pattern recognition, abstraction, algorithm design, and scale (Wing, 2006).

The DCIC at the University of Maryland's iSchool, in collaboration with all the authors of this chapter, is in the process of testing elements of CAS through interdisciplinary research themes that aim to help gain new digital skills, conduct interdisciplinary research, and explore professional development opportunities at the intersection of archives, big data, and analytics. This is grounded in projects that leverage unique archival collections in the following areas: refugee narratives, community displacement, racial zoning, movement of people, citizen internment, and cyberinfrastructure for digital curation.

Further development of the CAS agenda will require the development of transdisciplinary iSchools with faculty from computer science, archival science, and data science. To successfully inject the contributions of these different disciplines into courses will require collaborative development of syllabi and team teaching. Ensuring of students master basic skills, but at the same time learning of how to think flexibly to adapt to rapidly changing technological environments in which records are created and used, will require extensive hands-on experience working with cyberinfrastructure to carry out archival functions. Development of a CAS curriculum will require cross-disciplinary collaboration at levels far beyond those usually found in iSchools.

REFERENCES

Ambacher, B. (Ed.). (2003). *Thirty years of electronic records.* Lanham, MD: Scarecrow Press.

Bailey, S., & Vidyarthi, J. (2010). Human–computer interaction: The missing piece of the records management puzzle? *Records Management Journal*, 20(3), 279–290.

Beagrie, N. (2006). Digital curation for science, digital libraries, and individuals. *International Journal of Digital Curation, 1*(1), 3–16.

Blanke, T., Bryant, M., & Hedges, M. (2013). Back to our data: Experiments with NoSQL technologies in the humanities. *Big Data, 2013 IEEE International Conference on Big Data* (pp. 17–20).

Brohan, P., Allan, R., Freeman, J. E., Waple, A. M., Wheeler, D., Wilkinson, C., & Woodruff, S. (2009). Marine observations of old weather. *Bulletin of the American Meteorological Society*, *90*(2), 219–230.

Burgess, L. C. (2016). Provenance in digital libraries: Source, context, value and trust. In V. Lemieux (Ed.), *Building trust in information* (pp. 81–91). Geneva: Springer International Publishing.

Clark, D., McKeon, R., Marciano, R., & Bailey, M. (1998). Rear-projecting virtual data onto physical terrain: An exercise in two senses being better than one. In *Proceedings of IEEE Visualization.*

Cunningham, A. (2016). Describing archives in context: Peter J Scott and the Australian "series" system. In V. Lemieux (Ed.), *Building trust in information: Perspectives on the frontiers of provenance* (pp. 49–58). Berlin: Springer-Verlag, Research Triangle Park, NC.

Dingwall, G., Marciano, R., Moore, R., & McLellan, E. (2007). From data to records: Preserving the geographic information system of the city of Vancouver. *Archivaria, 64*, 181–198.

Dunn, S., & Hedges. M. (2013). Crowd-sourcing as a component of humanities research infrastructures. *International Journal of Humanities and Arts Computing*, *7*(1–2), 147–169.

Duranti, L., & Michetti, G. (2015). The archival method: Rediscovering a research tradition. In A. Gilliland, A. Lau, & A. McKemmish (Eds.), *Research in the archival multiverse.* Clayton, Victoria, Australia: Monash University Publishing.

Esteva, M., Sweat, S., McLay, R., Xu, W., & Kulasekaran, S. (2016). Data curation with a focus on reuse. In *Proceedings of the 2016 IEEE Joint Conference on Digital Libraries*, Newark, NJ: ACM Digital Library. Retrieved from http://dx.doi.org/10.1145/2910896.2910906

Fleet, C., Kowal, K., & Pridal, P. (2012). Georeferencer: Crowdsourced georeferencing for map library collections. *D-Lib Magazine, 18*(11/12). Retrieved from http://www.dlib.org/dlib/november12/fleet/11fleet.html

Heard, J., & Marciano, R. (2011). A system for scalable visualization of geographic archival records. In *Proceedings of the 1st IEEE Symposium on Large-Scale Data Analysis and Visualization* (pp. 121–122). Retrieved from http://www.slideshare.net/richardjmarciano/a-system-for-scalable-visualization-of-geographic-archival-records/

Jansen, G., Marciano, R., Padhy, S., & McHenry, K. (2016). Designing scalable cyberinfrastructure for metadata extraction in billion-record archives. In *iPRES2016*, Bern, Switzerland.

Lemieux, V. (2012). Using information visualization and visual analytics to achieve a more sustainable future for archives: A survey and critical analysis of some developments. *Comma, 2012*(2), 55–70.

Lemieux, V. (2014). Towards a 'third order' archival interface: Research notes on some theoretical and practical implications of visual explorations in the Canadian context of financial electronic records. *Archivaria, 78*, 53–93.

Lemieux, V. (2016a). *Building trust in information: Perspectives on the frontiers of provenance.* Berlin: Springer-Verlag.

Lemieux, V. (2016b). *Blockchain for recordkeeping: Help or hype?* Ottawa: Social Sciences and Humanities Research Council of Canada.

Lemieux, V. (2016c). Trusting records: Is blockchain technology the answer? *Records Management Journal, 26*(2), 110–139.

Ludäscher, B. (2016). A brief tour through provenance in scientific workflows and databases. In V. Lemieux (Ed.), *Building trust in information: Perspectives on the frontiers of provenance* (pp. 103–126). Berlin: Springer-Verlag.

Marciano, R. (2015). Revisiting inequality through the lens of crowdsourcing [Keynote address]. In *Citizen humanities comes of age: Crowdsourcing for the humanities in the 21st century.* Retrieved from https://connected-communities.org/index.php/news/citizen-humanitiescomes-of-age-crowdsourcing-for-the-humanities-in-the-21st-century-event-summary

Marzullo, K. (2016). *The federal big data research and development strategic plan*. Washington, DC: NITRD. Retrieved from https://www.nitrd.gov/PUBS/bigdatardstrategicplan.pdf

Molloy, L., Gow, A., & Konstantelos, L. (2014). The DigCurV curriculum framework for digital curation in the cultural heritage sector. *IJDC*, 9(1), 231–241.

Networking and Information Technology Research and Development Program. (2016). Supplement to the President's Budget FY 2017. Retrieved from https://www.nitrd.gov/pubs/2017supplement/FY2017NITRDSupplement.pdf

Northwestern Libraries. (2017). Data management: Federal funding agency requirements, a guide to assist researchers in data management. Retrieved from http://libguides.northwestern.edu/datamanagement/federalfundingagency

Panama Papers. (2016). Retrieved from https://panamapapers.icij.org

Pearce-Moses, R. (2005). *A Glossary of Archival and Records Terminology*. Chicago, IL: Society of American Archivists. Retrieved from http://www.archivists.org/glossary/index.asp

Software Prototyping. (2016). Retrieved from https://en.wikipedia.org/wiki/Software_prototyping/

Thibodeau, K. (2016). Research issues in archival provenance. In V. Lemieux (Ed.), *Building trust in information: Perspectives on the frontiers of provenance* (pp. 69–80). Berlin: Springer-Verlag.

Travis, D., Lee, M., Rojas, M., Gunn, A., Nimkar, A., Jansen, … Marciano, R. (2016). Unlocking the archives of displacement and trauma: Revealing hidden patterns and exploring new modes of public access through innovation and infrastructure. In *Archiving Conference* (pp. 135–139). Washington, DC: Society for Imaging Science and Technology.

Underwood, W. (1999). *Analysis of presidential electronic records: Final report*. Atlanta, GA: Georgia Tech Research Institute. Retrieved from http://perpos.gtri.gatech.edu/publications/PERPOS%20TR%201999-01.pdf

Underwood, W. (2002). A formal method for analyzing the authenticity properties of procedures for preserving digital records. In *Proceedings of the International Conference on Digital Archive Technologies*, pp. 53–64. Retrieved from http://www.iis.sinica.edu.tw/APEC02/Program/William.Underwood.doc

Underwood, W. (2005). *Mars global surveyor data records in the Planetary Data System: A case study*, InterPARES Technical Report. Atlanta, GA: Georgia Tech Research Institute.

Underwood, W. (2010). Grammar-based recognition of documentary forms and extraction of metadata. *International Journal of Digital Curation*, 5(1), 148–159. Retrieved from http://www.ijdc.net/index.php/ijdc/article/viewFile/152/215

Underwood, W., Kindl, M., Underwood, M., & Laib. S. (2001). *Presidential electronic records PilOt system (PERPOS): Phase I report.* Atlanta, GA: Georgia Tech Research Institute. Retrieved from http://perpos.gtri.gatech.edu/publications/PERPOS%20TR%202001-01.pdf

Underwood, W., & Laib, S. (2012). *Automatic categorization of email for records retention. Technical report*. Atlanta, GA: Information and Communications Laboratory, Georgia Tech Research Institute.

Underwood, W., Simpson, R., Whitaker, E., Iwanska, L. Laib, S., Harris, B., …, & Kau, J. (2006). *The presidential electronic record pilot system (PERPOS): Phase II.* Atlanta, GA: Georgia Tech Research Institute. Retrieved from http://perpos.gtri.gatech.edu/publications/PERPOS_TR%2006-06_%20Final_Scientific_Technical_Report.pdf

Upward, F. (2005). The records continuum. In S. McKemmish, M. Piggott, B. Reed, & F. Upward (Eds), *Archives: Recordkeeping in society* (pp. 197–222). Wagga Wagga, NSW, Australia: Centre for Information Studies.

Weijia, X., Esteva, M., Jain, S. D., & Jain, V. (2011). Analysis of large digital collections with interactive visualization. Austin, TX: University of Texas. Retrieved from https://www.cs.utexas.edu/~suyog/vast.pdf

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

Winling, L., Connolly, N., Nelson, R., & Marciano R. (2017). Integrating the Depression-era history of race and redlining. In I. Gregory & D. Lafreniere (Eds.), *Routledge handbook of spatial history*, (pp. 502–524). New York, NY: Routledge.

Zarro, M., & Hall, C. (2012). Exploring social curation. *D-Lib Magazine*, *18*(11/12). Retrieved from http://www.dlib.org/dlib/november12/zarro/11zarro.html