Stirring The Cauldron: Redefining Computational Archival Science (CAS) For The Big Data Domain

Nathaniel Payne
School Of Library, Archival, and Information Studies (iSchool)
University Of British Columbia
Vancouver, Canada
njpayne@mail.ubc.ca

Abstract— **Over the past 10 years, digitization, big data, and technology advancement has had a significant impact on the work done by computer scientists, information scientists, and archivists. Together, each of these groups has contributed to unlock new areas of trans-disciplinary research that are critical for forward progression in the world of big data, while collectively spurring the creation of a new inter-disciplinary field – Computational Archival Science (CAS). Unfortunately, significant gaps exist, including the lack of a comprehensive definition of CAS. This paper closes those gaps by proposing a new, comprehensive definition of Computational Archival Science (CAS) while simultaneously highlighting key big data challenges that exist both in industry and academia. The paper also proposes important areas of future research especially in the context of big data and artificial intelligence.**

Keywords—big data, computational archival science, provenance, computational science, transdisciplinarity, machine learning, artificial intelligence

I. Introduction

Over the past 10 years, digitization, big data, and technology advancement has had a significant impact on the work done by computer scientists, information scientists, and archivists [1][2]. Together, each of these groups has contributed to and challenged our understanding of the role that the disciplines of computer science, information science, and archival science play in the world of big data [3]. More importantly, each of these disciplines has individually contributed to unlock new areas of trans-disciplinary research that are critical for forward progression in the world of big data, while collectively spurring the creation of a new inter-disciplinary field – Computational Archival Science (CAS) [4]. While this new field offers great promise, in the words of Manfred Max-Neef, the movement " is still in the making"[5]. Significant gaps exist, including the lack of a collective, detailed research framework, which is essential for focused progress against many of the most challenging multi-disciplinary big data problems facing industry and academia today. What's more, due to an argued lack of general understanding of the approach that each of the individual disciplines take when it comes to the studying of information and data, the creation of a coherent research agenda which directly addresses the emerging opportunities in science, engineering, medicine, healthcare, and business, is essential.

Thus, in an effort to address the gap and lay a foundation for research, this paper will seek to explore the role that the disciplines of computational science, information science, and archival science play in this new interdisciplinary field of Computational Archival Science (CAS). More importantly, this paper will start to identify research areas and approaches that will transform how professional and research groups approach big data management, search and mining, security, privacy, and trust. While this is indeed a difficult task, it is no doubt a critically important task that is necessary due to the new and emerging digital document and information forms that continue to challenge all areas of the big data research spectrum [6][7][8]. This clarity is especially urgent because of the fact that the notion of what a record is, what an archive is, what information is, and even what knowledge is, continues to become more complex in this new digital age [6][9]. Resolution, thus calls for a much more rigorous evaluation of the roles that each of the disciplines will play in a combined future collaborative research framework. Through this evaluation, clarity will be gained on the areas of deficiency within the current definition of Archival Computational Science (CAS), the opportunities that exist from a practical perspective, and the critical points of future work that are needed to enable the acceleration of current and future research work.

II. Starting Point: Defining Computational Archival Science (CAS)

Computational Archival Science (CAS) is currently defined as:

> *"An interdisciplinary field concerned with the application of computational methods and resources to large-scale records/archives processing, analysis, storage, long-term preservation, and access, with 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 material."* [4].

This definition was updated by Marciano et al (2018) with the word "transdisciplinary":

> *"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."* [10]

As is clear, CAS is a multi-disciplinary field that is designed to reflect emerging challenges that exist both within academia and industry. It is a field that has many influences, continues to evolve, but has some key conceptual foundations that are critical to the three domains underlying it. One of these concepts which is critical, and which holds the key to understanding the contributions that each of the disciplines must make in the emerging big data world is "provenance". As noted by the International Council on Archives, provenance can be defined as:

> *"The relationships between records and the organizations or individuals that created, accumulated, and/or maintained and used them in the conduct of personal or corporate activity. Provenance is also the relationship between records and the functions which generated the need of the records."* [11],[4].

Provenance, while misunderstood, is deeply important to each of the three disciplines discussed in this paper, with the study of provenance itself presenting a rich field for exploration in the big data world particularly given the range of open research challenges that exist. Importantly, from a bridging perspective, each of the three disciplines that form the new foundational field of CAS look at and consider provenance very differently. This difference, both prospectively and retrospectively, can be attributed to the differing core fundamental lenses that each of the disciplines takes when approaching problems faced. These differences are key to understanding the importance of transdisciplinarity to making progress within the world of big data.

III. Understanding Transdisciplinarity

Before diving into an analysis of the three disciplines that are both central to the evolution of CAS and arguably the entire big data research agenda, it is important to have a solid framework against which to compare the disciplines against. Thus, to explore the question of transdisciplinarity, Max-Neef's model of disciplinary evaluation has been selected. This model provides a foundational 4 question framework that can be used to understand and evaluate each discipline and create a unifying framework for future work and new CAS definition. When utilizing this model, there are four main questions that researchers are encouraged to explore when evaluating a discipline [5]:

- **Question 1:** What must a discipline do? / "Must Do" - "How does what we propose to do contribute to understanding or doing what we must do, as a matter of values and ethics?"
- **Question 2**: What does a discipline want to do? / "Want To Do": "How does what we propose to do contribute to understanding or doing what we want to do in support of what we must do?"
- **Question 3**: What can a discipline do? / "Can Do": "Can we do what we must do and want to do?"
- **Question 4**: What can a discipline know? / "Can Know": "What can we know about what we propose to do?"

In exploring Max-Neef's work, one must pay special attention to Max-Neef's own arguments around the lack of connectedness between many disciplines. As Max-Neef argues, strong transdisciplinarity in most disciplines is still in the making [5]. Indeed, when trying to understand the transdisciplinary nature of a new field like CAS, it is not as simple as just worrying about how to orient traditional archival studies to new and emerging digital document and information forms. This is because the notion of what a document represents and of how archives create and sustain public or collective memory are continuing to evolve [4],[6],[39],[41]. Thus, in order to create a truly transdisciplinary research agenda, unlocking solutions to current problems within the big data domain, and accelerating the discovery of new solutions that will change the way that industry and academia work within the area of big data, we begin by independently reviewing the three disciplines that make up the foundation of computational archival science, while then working to synthesize a unifying framework and create a singular definition that can be used as the foundation for a comprehensive, forward looking research agenda.

A. Understanding Archival Science

We begin our first analysis of the three disciplines with a review of archival science. Archival science is the academic and professional discipline concerned with the theory, methodology, and practice of the creation, preservation, and use of records and archives [41]. Archival science encompasses the creation, preservation, and use of records in their functional context, whether organizational or personal, and the wider social, legal, and cultural environment within which records are created and used. Within the discipline of archival science, the central problem is to ensure that records are persistent in time while also ensuring that records remain as special representations of things over time [12]. This means that the most important item to study is the archival document, also known as the record. As Duranti notes, an archival document is a document created or received by of physical or juridical person in the course of practical activity and preserved [42]. Archival documents are defined by their archival nature. In this sense, the archival nature refers to the whole of the characteristics with which archival documents are endowed by the circumstances of their creation and which are therefore natural to them. Those characteristics are authenticity, impartiality, interrelatedness, naturalness, and uniqueness [43].

In general, archivists and those working within archival science domain, want to ensure that the representations of records have longevity [4]. They are focused heavily on context and the archival bond [44]. This is because records cannot be fully understood without adequate knowledge of the activity which gave rise to them, the wider function of which that activity forms part, and the administrative context, including the identities and roles of the various participants in the activity [41]. Thus, contextual information must be captured in the records themselves or in the systems that are used to maintain them.

In addition to context, authenticity and trustworthiness are critically important within the archival science domain. Records, must have the qualities of authenticity, integrity, usability, and reliability [45]. Thus, authenticity and integrity of records need to be guaranteed over time so that users can be confident that records are genuine and trustworthy and that no illicit alterations have been made to them. Once these qualities are established, archival science as a discipline is then concerned with using these artefacts and qualities to represent a fact that relates to an act and which exists between two or more parties. In this case, the record becomes that representation of the transaction. With this knowledge, archival science researchers and practitioners can then know authenticity [46].

B. *Understanding Information Science*

While archival science focuses heavily on the record and the archive, information science as a discipline takes a differing and very important focus - the human. Information science is the science and practice dealing with the effective collection, storage, retrieval, and use of information. It is concerned with recordable information and knowledge, and the technologies and related services that facilitate their management and use. [59]. More specifically, information science is a field of professional practice and scientific inquiry addressing the effective communication of information and information objects, particularly knowledge records, among humans in the context of social, organizational, and individual need for and use of information [59],[60].

From a domain perspective, the domain of information science is the transmission of the universe of human knowledge in recorded form, centering on manipulation (representation, organization, and retrieval) of information, rather than knowing information [61]. Information science often views information networks as socio-technical constructs, taking a particularly human first focus. This, without surprise, is because of the two key orientations of the discipline: 1) Toward the human and social need for and use of information pertaining to knowledge records 2) Toward specific information techniques, systems, and technologies (covered under the name of information retrieval) to satisfy that need and provide for effective organization and retrieval of information. This creates two disparate orientations for the discipline, one that deals with information need, or more broadly human information behavior, and the other that deals with information retrieval techniques and systems. These two orientations are themselves the foundation for the intellectual framework for the discipline, which Bates broke into three distinct questions that still remain relevant for today.

- The *physical question*: What are the features and laws of the recorded information universe?
- The *social question*: How do people relate to, seek,
- and use information?
- The *design question*: How can access to recorded information be made most rapid and effective?

From a practice perspective, information scientists are generally focused on understanding information communities that build up around systems and technologies, while also understanding the information behaviors that occur in a variety of settings. Information scientists want to deeply understand the human aspect of information and technology interaction while also understanding how humans interact with information, how they use it, and how they access it [15][16]. With this approach, information scientists then seek to better understand the information behaviors that exist in a wide variety of settings [17]. They then use this understanding to know how human actors process information in particular systems and thus how to optimize their experience and interaction [18].

Without question, the perspective of the information scientist and the discipline as a whole is absolutely critical. Unfortunately, it is a perspective that is missing from the current discussions around CAS due to the strong archival and computational focus and is an area that is critically important in a larger discussion of big data research. For example, advancing research around the area of social web search and mining relies on an understanding of the communities and humans that impact the social web [62],[63]. Indeed, humans and communities create meta-data that are important for analysis and linking, enabling the development of robust computational models that can be used to build large scale recommender systems and social media systems [64]. Without a strong focus on the human actor and the role of the community, it is very difficult to create precise and accurate machine learning models of communities [65]. Such accuracy and precision is critical within domains such as public health, where individual meta-data can help one determine important predictors for disease or help find other anomalies [66],[67]. Indeed, the strength of a graph network can be argued to be related to the strength of the linkages between the various nodes [68]. Without the information science lenses, these critical problems which are important to both big data research and the wider CAS domain will not move forward efficiently. This lens, is also critical to the important research going on within the big data community around human computer interaction. Understanding deeply how people use, access, and process information, as well as methods that relate to things like information foraging when information is distributed, are essential if we hope to build systems that maximize the contribution of the human actor and improve the engagement of humans with technology [69], [19], [14].

C. *Understanding Computer Science*

As opposed to both archival science and information science, computer science and computer scientists take a markedly different perspective. As Denning notes, computer science is the body of knowledge dealing with the design, analysis, implementation, efficiency, and application of processes that transform information. The fundamental question underlying all of computer science is what can be automated [70]. The single most central question, according to Rapaport and others, is what can be computed, and how [71]? From this starting point, four additional questions following logically that frame the computer science perspective: What can be computed *efficiently*, and how? What can be computed *practically*, and how? What can be computed *physically*, and how? What *should* be computed, and how [71]?

With the focus on computation, it is not difficult to see why computer science is often called the study of algorithms, and more broadly, the science of computation and algorithms. [72],[73],[74]. Some definitions have substituted computers for computation, since, as is argued, one needs computers in order to properly study algorithms because human beings are not precise enough nor fast enough to carry out any but the simplest procedures [75]. This is particularly true in areas like deep learning where we need computers in order to understand and test whether "deep learning" algorithms really do what they are intended to do, and do so in real time [75].

Finally, while there is large agreement on the focus on computation and algorithms, there continues to be an evolving focus on information. Information, as Samuel Johnson initially defined, can be referred to as intelligence given [76]. Others, including Duranti, have referred to information as a message or knowledge which has been voluntarily conveyed, or a message intended for communication over time and space [77],[78]. Accepting these definitions of information, we can then see how information can be argued to be a central focus of computer science. As Forsythe noted originally, computer science is not the study of computers or of algorithms, but of information [79]. Others have agreed, including Denning, who notes that, at its foundation, computer science is the art and science of representing and processing information and, in particular, processing information with the logical engines called automatic digital computers [80],[81]. This focus was embedded in Denning's extended discussion on computer science where he identified the fundamental question, suggesting that computer science, at its core, is simply the body of knowledge dealing with the design, analysis, implementation, efficiency, and application of processes that transform information [70]. From an information perspective, computer science studies how to represent and (algorithmically) process information, as well as the machines and systems that do this. As Hartmanis and Lin note elegantly [82]:

> *For the physicist, the object of study may be an atom or a star. For the biologist, it may be a cell or a plant. But computer scientists and engineers focus on information, on the ways of representing and processing information, and on the machines and systems that perform these tasks.*

Looked at collectively, the various defining perspectives on computer science provide us a stable foundation for discussions relating to the discipline of computational archival science. At its core, one can conclude that, rather than being focused on the record from an archival science perspective, or on humans from an information science perspective, computer scientists focus generally on the systems that are being created and the various computational systems that are being used for a variety of purposes. This grounds the discipline both theoretically and from an applied perspective in computation, makes computer science an essential element of Computational Archival Science [20]. While there are many different research areas that computer science pursues, its focus on the feasibility, structure, expression, and mechanization of algorithms, systems, and networks, as well as critical research into how to most effectively acquire, represent, process, store, communicate, and access information, highlight the central driving role that computer scientists have to play in the long term CAS research agenda.

Over the last few years there has been an increasing blending between computational science and archival science, especially in the area of big data. This blending is perhaps most noticeable in the areas of provenance. To the computer scientist, provenance is important from a records perspective, particularly when looking at systems, artefacts, individual records within systems and their processing. Just as an archival scientist wants to understand how a record is shaped over time, the computational scientists has many reasons to want to understand this same information. Additionally, the computational scientist is also focused clearly on understanding how this provenance changes over a time window and the impact of this change on systems, performance, algorithms, etc. Such research is critical within the discussion of big data, especially within the area of big data security, privacy and trust.

IV. Seeing Three Disciplines As One – Building A Foundation For Big Data and CAS

In order to bring the disciplines together and start to understand the transdisciplinary opportunities that exist for both the CAS domain and big data research environment, we begin with an inversion exercise that aims to blend together the key attributes from Max-Neef's analysis for each of the disciplines. By articulating the key elements of each discipline in this way, one can then compare the similarities and differences between the various fields and more effectively draw conclusions. The list of items is shown in Figure 1.

	Disciplines		
	Archival Science	***Information Science***	***Computational Science***
Must Do	Understand how we make records persistent in time	Understand the human aspect of information and technology interaction	Understand the theory of computation and optimal design of a system
Want To Do	Focus on the record and presrve it	Understand the socio-technical construct of information networks	Apply this understanding to problems while enabling best practice design and computation
Can Do	Understand the facts that relate to an act of transaction	Understand the information behaviors of humans interacting in a wide variety of settings	Understand the feasibility, structure, expression, and mechanization of algorithms, systems, and networks
Can Know	Authenticity and truthfulness	Know how human actors process information and optimize their	How to most effectively acquire, represent, process, store, communicate,

	Disciplines		
	Archival Science	*Information Science*	*Computational Science*
		experience and interaction	and access information

Fig. 1. Seeing The Disciplines Through A Multi-Disciplinary Lenses

As we reflect on the table, the difference in the focus, goals, and approaches starts to become clear. These differences relate to the state of the CAS discipline today and indeed much of the research going on within the big data research domain in which weak coupling exists between the knowledge in each of these disciplines. Indeed, as is common in big data practice, a person may have studied, simultaneously or in sequence, more than one area of knowledge, without making any connections between them. One may, for example, become competent in archival science, information science, or computational science, without generating any cooperation between the disciplines. What's more, while research intent of often transdisciplinary, it is arguable that, especially within the domain of CAS and big data, research practice is multi-disciplinary at best, with many multidisciplinary teams of researchers and technicians carrying out analysis and research separately from each other and separate from implementation. The end results of these collaborations are often seen from the perspective of their individual disciplines, with the final result being a series of "reports pasted together, without any integrating synthesis." [5]

In order to start closing these gaps particularly as it relates to CAS, we begin by making the decision to pivot our thinking and perspective and removing the disciplinary focus. By doing so, we move closer to a more integrated understanding and definition of computational archival science, while highlighting the key themes discovered in our research. This is shown in Figure 2.

Computational Archival Science (CAS)			
Must Do	*Want To Do*	*Can Do*	*Can Know*
Record persistent in time and space	Preserve the record	Facts related to the act of transaction	Authenticity and truthfulness
Human aspect of information and technology interaction	Socio-technical construct of information networks	Information behaviors of humans interacting in a wide variety of settings	Human actor information processing information and optimize their experience and interaction
Understand the theory of computation and optimal system design	Optimized practice design and computation	Optimized feasibility, structure, expression, and mechanization of algorithms, systems, and networks	Optimal information acquisition, representation, processing, storage, communication, and access

Fig 2. Seeing The Disciplines Through A Single-Disciplinary Lenses

As can be seen in figure 2, by changing our own lenses, the archival, information, and computer sciences themes are all emphasized within the one multi-disciplinary construct. For example, working together as a single discipline, the archival scientists focus on understanding how to make records persistent over time is balanced against the information scientists desire to understand the human aspect of information and technology interaction. This, in turn, is balanced against the computer scientists desire to understand the theory of computation & optimal design of systems. This balance is absolutely essential for solving some of the most pressing big data problems that we are facing today, and is important when one seeks to create a single, comprehensive definition of computational archival science.

A. Analyzing The Components

Now that we understand the base elements that could make up the CAS field in a multi-disciplinary or "pluridisciplinarity" approach, we then turn to analyzing the components against the initially proposed definition of CAS. As was originally noted, the initial definition of CAS was:

> *"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."* [10]

As is evident in the above, the must do items from archival science are well covered within the initial discipline, with a focus on long term preservation, arrangement and description being well identified. From an information science perspective, the only words that are referenced within the definition that relate to the information science perspective are "access" and "access decisions". This highlights that only weak reference to the information science domains is incorporated within the current definition. Indeed, no reference is made to the human operator, or any human focused technology impact,. This is a significant weakness, especially in the context of big data. Finally, we see that the current definition contains major gaps from a computational science perspective. From a must do perspective, the definition of CAS does refer to the "application of computational methods and resources". That said, it is a question whether this also refers to any theoretical research or computational science approaches. There is also no comment or reference relating to understanding the best practice computational and system design, a critical problem domain within computer science. This strongly applied perspective leaves much room for future work and is one of the many findings from this initial research exploration.

Moving forward, we then shift our focus to the "Can Do" and "Can Know" dimensions of Max-Neef's framework [5]. In doing this, we see that there are major gaps that relate to the current definition. For example, there is no reference to understanding facts that relate to an act or transaction in a specific way other than an implied relationship to the core archival forms of description. There is also no specific comment referencing authenticity, truthfulness, or the language used commonly within the field of diplomatics.

With this in mind, it is clear that there are significant opportunities for future work and the evolution of the current definition of CAS, including more research time spent understanding how the concepts of authenticity and truthfulness will be reflected in the CAS and the big data domain. There is also no clear link between the aims of the initial definition, which including improving efficiency, productivity, and precision and the core disciplines. Is pursuing efficiency, for example, purely a computer scientific pursuit that relates to workflows, or a human centric approach that needs the input of an information science perspective. Both, arguably, are needed, especially as one considers the ongoing changes [87].

In order to work toward a final unifying definition and framework for CAS, we now turn back and revisit the initial layout from our model from Figure 2, which shows, in italics and bold, the deficiencies and opportunities that exist for research collaboration and growth. This is shown in Figure 3.

Computational Archival Science (CAS)			
Must Do	***Want To Do***	***Can Do***	***Can Know***
Record persistent in time and space	Preserve the record	***Facts related to the act of transaction***	***Authenticity and truthfulness***
Human aspect of information and technology interaction	***Socio-technical construct of information networks***	***Information behaviors of humans interacting in a wide variety of settings***	Human actor information processing information and optimize their experience and interaction
Understand the theory of computation and optimal system design	***Optimized practice design and computation***	***Optimized feasibility, structure, expression, and mechanization of algorithms, systems, and networks***	Optimal information acquisition, representation, processing, storage, communication, and access

Fig 3. Understanding The Gaps

As is shown in Figure 3, while this new inter-disciplinary field is well on its way to a forming, clear gaps within the want to do and can do areas pose limitations and create opportunities for future research. These gaps also motivate the need for a new definition of Computational Archival Science (CAS) which is inclusive, transdisciplinary, and forward looking. As is proposed, there are 5 key elements of this new definition. These elements see Computational Archival Science (CAS) defined as:

- ***A transdisciplinary field grounded in archival, information, and computational science that is …***
- ***concerned with the application of computational methods and resources, design patterns, socio-technical constructs, and human-technology interaction,***
- ***to large-scale (big data) records/archives processing, analysis, storage, long-term preservation, and access problems,***
- ***with the aim of improving and optimizing efficiency, authenticity, truthfulness, provenance, productivity, computation, information structure and design, precision, and human technology interaction***
- ***in support of acquisition, appraisal, arrangement and description, preservation, communication, transmission, analysis, and access decisions***

Said together, the new definition of Computational Archival Science can be stated as the following:

> *Computational Archival Science (CAS) is a transdisciplinary field grounded in archival, information, and computational science that is concerned with the application of computational methods and resources, design patterns, socio-technical constructs, and human-technology interaction to large-scale (big data) records/archives processing, analysis, storage, long-term preservation, and access problems with the aim of improving and optimizing efficiency, authenticity, truthfulness, provenance, productivity, computation, information structure and design, precision, and human technology interaction in support of acquisition, appraisal, arrangement and description, preservation, communication, transmission, analysis, and access decisions.*

As one stops to reflect on this new, more comprehensive definition, it is useful to review this definition in light of some of the areas within the big data research world where a new approach - which truly incorporates all disciplinary perspectives from archival, information, and computational science background - appears fruitful. For example, while researchers like Avison & Elliot have proposed developing new theory for big data problems relating to optimal computational & system design for distributed systems, theoretical and practical work in these areas has not moved forward due to a lack of proper consideration of provenance – a core archival science construct [26]. From the outside, one would assume that distributed systems represent a fruitful area for future research especially within the area of big data and CAS. Within the distributed systems landscape, understanding both retrospective and prospective provenance can provide great benefit to individuals working with such systems, developing workflows, conducing and development new methods for data audit, and many others. That said, as Dr. Lemieux points out, distributed systems make it challenging to capture provenance from "processes that are distributed over multiple, heterogeneous, autonomous systems. Each of these systems may be expected to provide some fragment of provenance, requiring post hoc composition of these fragments." [4].

What's more, as vast networks of interconnected information and processing systems are put into place, storage and retrieval are bound to be issues that also deserve research attention. Again, these areas, as well as other, remain underserved. This is surprising, especially considering the big data challenges that exist and which are impacted by this work in the areas of social web search and mining, peer-to-peer search, cloud, grid, and stream data mining, as well as link and graph mining. While there are various arguments that one could propose around why these gaps exist, our exploration of the initial deficits highlights the need to form "a deeper understanding of provenance itself is needed in order to cope with new forms of documentation and new modes of communicating and processing information." [4] For example from an archival perspective, one critical outstanding issue will require us to solve the problem of identifying who can be considered the creator of an archival object. This is particular true as organizations change at an ever increasing rate [4].

V. Exploring Emerging Problem Domains & Missed Opportunities

Now that we have completed our analysis of the current definition and proposal of a new comprehensive definition, we turn our attention to focusing on the evolving research areas that can benefit especially within the context of big data. In this quick discussion, we will seek to understand the opportunities for future work which can be covered using a comprehensive research agenda.

Over the past couple of years, CAS researchers have started focusing their energy on a number of key areas, including:

- Archival material analysis including text-mining, data-mining, sentiment analysis, network analysis.
- Scalable services for archives and archival processing, including identification, preservation, metadata generation, integrity checking, normalization, reconciliation, linked data, entity extraction, anonymization and reduction. Archival here includes appraisal, arrangement and description.
- Development of new forms of archives, including Web, social media, audiovisual archives, and blockchain.
- Cyber-infrastructures for archive-based research and for development and hosting of collections.
- Big data and archival theory and practice. This includes digital curation and preservation. Crowd-sourcing and archives. Big data and the construction of memory and identity. Specific big data technologies (e.g. NoSQL databases) and their applications. Corpora and reference collections of big archival data. Linked data and archives. Big data and provenance. Constructing big data research objects from archive.

While these are excellent starting point, a comprehensive review of the literature and the problem areas within big data shows many areas where significant opportunities for CAS related research existed that had been previously but indirectly flagged by researchers in other fields. These focus areas for future research include:

- Machine learning, prediction & forecasting research [27], including research relating to deep learning methods and other statistical methods, as well as the optimal design of algorithms, that can correctly classify and categorize records and their resulting meta-data [88],[89],[90],[91],[92],[93].
- Natural language understanding research which will transform our current, primitive, AI text analysis capabilities and truly understand the context of language. Such research is critical to enabling us to build machines that can truly interact with us. [94],[95],[96],[97]
- High performance computing [28] research, including specific work in the areas algorithms, computability & complexity that is directly related to CAS [98],[99],[100].
- Human computer interaction (HCI) research that supports information systems work [29], including an understanding of the role of the human in autonomous technology operations settings [101],[102].
- Distributed ledger research including blockchain research that can explore how to optimally preserve the archival bond within database systems [103],[104],[105],[106].
- New methods for information accumulation, storage, search, and discovery, especially in information rich environments where multiple media are used as inputs for feature analysis and information retrieval.
- Detailed research into neuro-biology, especially research that enables a deeper understanding of how the human brain processes information.

In addition to this work, system design, architecture & information systems work that supports computational scientific needs relating to CAS [30] is needed. Research relating to operating systems may hold promise and enable us to break some of the linked data problems that we are facing, as well as problems in the areas of network & application security, software analysis & testing, computational vision, knowledge based artificial intelligence including reinforcement learning, computer networking research which will directly impact data provenance, robotics work, as well as education & educational technology related work.

From an applied perspective, there are also many specific fields that could benefit from the inclusion of the above into the wider CAS research body, including transportation & networks, financial services & banking, natural resources & geophysics [31], journalism, psychology & cognitive science [32], legal, crime & criminal justice [33], sociology and community research [34], digital transformation [35], enterprise risk management, data warehousing & database systems, and business technology management [36][37]. Future papers will be dedicated to exploring these areas in depth.

In addition to this, as noted by researchers including Dr. Lemieux, there remains a pressing need to develop solutions to "more easily extract provenance information" [4]. This is clearly pressing within the domain of information science, particularly for "human-in-the loop" cognitive systems that are designed to capture provenance from processes that are distributed over multiple, heterogeneous, autonomous systems (machine and human). The short time window for the capture of this information and the potential errors relating to the capture of this information are significant areas for future work because of their ability to negatively impact the capture of analytical provenance information. This judgement can also be clouded by the individuals own experience. Tackling this problem at scale will also require researchers from many disciplines to think about how to most effectively store, index, and retrieve the information.

Finally, as is not surprising, there are significant opportunities within the area of big data security research that are relevant and pressing. Within the areas of big data security, privacy and trust, intrusion and anomaly detection are critically important avenues that would benefit from a cross-disciplinary perspective, as is large scale network visualization. The development of methods that enable the location of personal and private information within large corpuses, methods that enable large scale information processing, especially visual and textual information, methods that enable more effective information search (supporting the challenges of eDiscovery and supervision), and other methods, would be very useful. There is also significant work that is needed relating to the methods used in large scale natural language processing, event prediction, big data search, autonomic computing, and records management.

VI. Conclusion

As is evident in this paper, while the current definition of CAS provided a great roadmap in in the past, the identified gaps, and newly proposed definition, provide a fruitful starting point that can open up significant opportunities for future work and collaboration in many areas. To succeed in building truly intelligence machines, we must start by using Computational Archival Science principles to build algorithms that can understand context, interact with human inputs, and store data and information in new ways. Success in this pursuit will be measured differently in both academic and industry, but will require significant work by many groups to bring together differing view point and various bodies of research work and practice while building unified CAS domain. As Max-Neef notes, strong transdisciplinarity is truly an unfinished project which demands many efforts of systematization to be undertaken. [5] In the case of big data and CAS, it is hard to dispute this argument, especially considering the problems that need to be addressed and the amazing opportunities that exist as we look ahead.

Acknowledgment

Special thanks to Dr. Victoria Lemieux, Associate Professor, Blockchain @UBC Cluster Lead, and my doctoral senior supervisor, for her continued support, inspiration, patience, and guidance.

References

[1] Benhamou E, Eisenberg J, Katz RH (2010) Assessing the changing U.S. IT R&D ecosystem. Communications ACM 53(2):76–83

[2] King JL, Lyytinen K, eds. (2006) Information Systems: The State of the Field (John Wiley & Sons, Chichester, UK).

[3] Dietrich, D. & Adelstein, F. (2015) Archival science, digital forensics, and new media art. Digital Investigation 14 (2015) S137-S145

[4] Lemieux, V. (Ed.) (2016) Building Trust in Information. Springer.

[5] Max-Neef, A. (2004). Foundations of transdisciplinarity. Ecological Economics 53; 5– 16

[6] Cox, R. & Larsen, R.L. (2008) iSchools and archival studies. Archival Science. 8; 307

[7] Benbasat I, Zmud RW (2003) The identity crisis within the IS discipline: Defining and communicating the discipline's core properties. MIS Quarterly 27(2):183–194

[8] Galliers, R. (2003) Change as crisis or growth? Toward a transdisciplinary view of information systems as a field of study: A response to Benbasat and Zmud's call for returning to the IT artifact. J. Assoc. Inform. Systems 4(1):337–351

[9] Herring, M. (2007) Fool's gold: why the Internet is no substitute for a library. McFarland, Jefferson

[10] 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

[11] Omitola, T.; Gibbins, N.; Shadbolt, N. (2010) Provenance in Linked Data Integration. Future Internet Assembly, Ghent, Belgium, 16-17 December.

[12] Cunningham A (2008) Digital curation/digital archiving: a view from the National Archives of Australia. Am Arch 71:530–543

[13] Pearce-Moses, R. (2005) A glossary of archival and records terminology. Society of American Archivists;

[14] Castro, G. & Costa, B. (2016). Using data provenance to improve software process enactment, monitoring and analysis. Proceeding. ICSE '16 Proceedings of the 38th International Conference on Software Engineering Companion. Pages 875-878. Austin, Texas — May 14 - 22, 2016.

[15] Bryant, A. (2008) The future of information systems—Thinking informatically. European Journal Of Information Systems. 17(6):695–698

[16] Goffman, W. (1970) Information science: Discipline or disappearance? Aslib Proc. 22(12):589–596

[17] Griffiths, J. (2000) Back to the future: Information science for the new millennium. Bull. Amer. Soc. Inform. Sci. 26(4):24–27.

[18] Hirschheim, R. & Klein, H. (2003) Crisis in the IS field? A critical reflection on the state of the discipline. J. Assoc. Inform. Systems 4(5):237–293

[19] Bearman, D., Lytle, R. (1985) The power of the principle of provenance. Archivaria. 1:21

[20] Denning, P. (2005) Is computer science science? Comm. ACM 48(4): 27–31

[21] Green, T.J., G. Karvounarakis, and V. Tannen, "Provenance semirings", in PODS'07, 2007, pp. 31–40

[22] Buneman, P., S. Khanna, W.-C. Tan (2002) "On propagation of deletions and annotations through views," in Proceedings of the 21st ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (PODS'02), pp. 150–158

[23] Buneman, P.; S. Khanna, and W. C. Tan (2001) "Why and where: A characterization of data provenance", in Proceedings of the 8th International Conference on Database Theory, pp. 316–330

[24] Oinn, M. Addis, J. Ferris, D. Marvin, M. Senger, M. Greenwood, T. Carver, K. Glover, M. R. Pocock, A. Wipat, and P. Li (2004). "Taverna: a tool for the composition and enactment of bioinformatics workflows", Bioinformatics, vol. 20, no. 17, pp. 3045–3054

[25] Sansrimahachai, W.; Moreau, L.; Weal, M. (2013) Supporting On-the-fly Provenance Tracking in Stream Processing Systems. International Journal of Computer & Information Science, Vol. 14, No. 2 , December 2013.

[26] Avison, D. & Elliot, S. (2006) Scoping the discipline of information systems. King JL, Lyytinen K, eds. Information Systems: The State of the Field (John Wiley & Sons, Chichester, UK), 3–18

[27] Cox, R. (2007) Machines in the archives: Technology and the coming transformation of archival reference. First Monday. 12(11-5); http://firstmonday.org/article/view/2029/1894

[28] Arms, W. (2008) Cyber scholarship: High Performance Computing Meets Digital Libraries. Journal of Electronic Publishing; 11(1)

[29] Yoo Y (2010) Computing in everyday life: A call for research on experiential computing. MIS Quart. 34(2):213–231.

[30] Grover, V. (2012) The information systems field: Making a case for maturity and contribution. J. Assoc. Inform. Systems 13(4)

[31] Bowker GC (2005) Memory practices in the sciences. MIT Press, Cambridge

[32] Bowker G. (1994) Science on the run: information management and industrial geophysics at Schlumberger, 1920–1940. MIT Press, Cambridge

[33] McKemmish, R. (1999) What is forensic computing? Trends Issues Crime. Criminal Justice; 118

[34] Cook T (2013) Evidence, memory, identity, and community: four shifting archival paradigms. Archival Science 13:95–120

[35] Avgerou, C. (2001) The significance of context in information system and organizational change. Information Systems Journal 11(1):43–63

[36] Hirschheim R. & Klein H. (2011) Setting the scene: Tracing the history of the information systems field.

[37] Hirschheim R, Klein HK (2012) A glorious and not so-short history of the information systems field. J. Assoc. Inform. Systems 13(4): 188–235.

[38] Woodruff, A. & Stonebraker, M. "Supporting fine-grained data lineage in a database visualization environment," in Proceedings of the 13th International Conference on Data Engineering, 1997, pp. 91–102.

[39] Duranti, L. (2001). Concepts, principles, and methods for the management of electronic records. The Information Society, 17(4), 271-279.

[40] Duranti, L. (2010). From digital diplomatics to digital records forensics. Archivaria, 68, 39-66.

[41] Shepherd, E. (2009). Archival science. In Encyclopedia of Library and information sciences (pp. 179-191). CRC Press.

[42] Duranti, L.(1998). Diplomatics: New Uses for an Old Science. Lanham, MD, and London: The Scarecrow Press. Archivaria 28. Part 1

[43] Duranti, L. (1994). The concept of appraisal and archival theory. The American Archivist, 57(2), 328-344.

[44] Duranti, L. (1997). The archival bond. Archives and Museum Informatics, 11(3-4), 213-218.

[45] Duranti, L. (1995). Reliability and authenticity: the concepts and their implications. Archivaria, 39.

[46] Duranti, L. (1998). Diplomatics: new uses for an old science. Scarecrow Press.

[47] Yeo, G. (2007). Concepts of record (1): evidence, information, and persistent representations. The American Archivist, 70(2), 315-343.

[48] Shepherd, E., & Yeo, G. (2003). Managing records: a handbook of principles and practice. Facet publishing.

[49] Yeo, G. (2008). Concepts of record (2): prototypes and boundary objects. The American Archivist, 71(1), 118-143.

[50] Yeo, G. (2011). Rising to the level of a record? Some thoughts on records and documents. Records Management Journal, 21(1), 8-27.

[51] Chen, S. S. (2007). Digital preservation: Organizational commitment, archival stability, and technological continuity. Journal of organizational computing and electronic commerce, 17(3), 205-215.

[52] Payne, N., & Baron, J. R. (2017, December). Auto-categorization methods for digital archives. In Big Data (Big Data), 2017 IEEE International Conference on (pp. 2288-2298). IEEE.

[53] Baron, J. R., & Payne, N. (2017, May). Dark Archives and Edemocracy: Strategies for Overcoming Access Barriers to the Public Record Archives of the Future. In E-Democracy and Open Government (CeDEM), 2017 Conference for (pp. 3-11). IEEE.

[54] Simmhan, Y. L., Plale, B., & Gannon, D. (2005). A survey of data provenance in e-science. ACM Sigmod Record, 34(3), 31-36.

[55] Lemieux, V. L. (2016). Provenance: Past, Present and Future in Interdisciplinary and Multidisciplinary Perspective. In Building Trust in Information (pp. 3-45). Springer, Cham.

[56] Moore, R., Rajasekar, A., & Marciano, R. (2007). Implementing trusted digital repositories. Retrieved December, 4, 2007.

[57] Duchein, M. (1983). Theoretical principles and practical problems of respect des fonds in Archival Science. Archivaria, 16, 64-82.

[58] Vicknair, C., Macias, M., Zhao, Z., Nan, X., Chen, Y., & Wilkins, D. (2010, April). A comparison of a graph database and a relational database: a data provenance perspective. In Proceedings of the 48th annual Southeast regional conference(p. 42). ACM.

[59] Saracevic, T. (2009). Information Science. In Encyclopedia of Library and information sciences (pp. 179-191). CRC Press

[60] Saracevic, T. (1999) Information science. J. Am. Soc. Info. Sci. 50 (12), 1051–1063.

[61] Bates, M. (1999) The invisible substrate of information science. Journal Of the American Society For Information Science. 1999, 50 (12), 1043–1050.

[62] Bao, S., Xue, G., Wu, X., Yu, Y., Fei, B., & Su, Z. (2007, May). Optimizing web search using social annotations. In Proceedings of the 16th international conference on World Wide Web (pp. 501-510). ACM.

[63] Heymann, P., Koutrika, G., & Garcia-Molina, H. (2008, February). Can social bookmarking improve web search?. In Proceedings of the 2008 International Conference on Web Search and Data Mining (pp. 195-206). ACM.

[64] Schafer, J. B., Frankowski, D., Herlocker, J., & Sen, S. (2007). Collaborative filtering recommender systems. In The adaptive web (pp. 291-324). Springer, Berlin, Heidelberg.

[65] Sun, N., Rau, P. P. L., & Ma, L. (2014). Understanding lurkers in online communities: A literature review. Computers in Human Behavior, 38, 110-117.

[66] Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.

[67] Panagiotakos, D. B., Dimopoulos, A. C., Caballero, F. F., & Haro, J. M. (2018). Machine Learning as an alternative of Statistical methods in predicting chronic disease risk. Annals of Epidemiology, 28(9), 658.

[68] Granovetter, M. (1983). The strength of weak ties: A network theory revisited. *Sociological theory*, 201-233.

[69] Orso, A., & Rothermel, G. (2014, May). Software testing: a research travelogue (2000–2014). In Proceedings of the on Future of Software Engineering (pp. 117-132). ACM.

[70] Denning, P. (1985). The Science of Computing: What is computer science? American Scientist, 73(1), 16-19.

[71] Rapaport, W. J. (2017). What is computer science. University at Buffalo, The State University.

[72] Knuth, D.E. (1974a). Computer programming as an art. Communications of the ACM, 17(12):667–673.

[73] Knuth, D.E. (1974b). Computer science and its relation to mathematics. American Mathematical Monthly, 81(4):323–343.

[74] Newell, A., Perlis, A.J., and Simon, H.A. (1967). Computer science. Science, 157(3795):1373–1374

[75] Lewis-Kraus, G. (2016). The great A.I. awakening. New York Times Magazine.

[76] Weller, T., & Bawden, D. (2006). Individual perceptions: a new chapter on Victorian information history. Library History, 22(2), 137-156.

[77] Zins, C. (2007). Conceptions of information science. Journal of the American Society for Information Science and Technology, 58(3), 335-350.

[78] Duranti, L. (1995). Reliability and authenticity: the concepts and their implications. Archivaria, 39.

[79] Forsythe, G.E. (1967). A university's educational program in computer science. Communications of the ACM, 10(1):3–80.

[80] Denning, P.J. (2007). Computing is a natural science. Communications of the ACM, 50(7):13–18.

[81] Denning, P.J. (2009). Beyond computational thinking. Communications of the ACM, 52(6):28–30.

[82] Hartmanis, J. and Lin, H. (1992). What is computer science and engineering? In Hartmanis, J. and Lin, H., editors, Computing the Future: A Broader Agenda for Computer Science and Engineering, pages 163–216. National Academy Press, Washington, DC. Ch. 6.

[83] Lim, C., Lu, S., Chebotko, A., & Fotouhi, F. (2010, July). Prospective and retrospective provenance collection in scientific workflow environments. In 2010 IEEE International Conference on Services Computing (pp. 449-456). IEEE.

[84] Davidson, S. B., & Freire, J. (2008, June). Provenance and scientific workflows: challenges and opportunities. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data (pp. 1345-1350). ACM.

[85] Freire, J., Koop, D., Santos, E., & Silva, C. T. (2008). Provenance for computational tasks: A survey. Computing in Science & Engineering, 10(3).

[86] Dey, S., Belhajjame, K., Koop, D., Raul, M., & Ludäscher, B. (2015). Linking prospective and retrospective provenance in scripts. Theory and Practice of Provenance (TaPP).

[87] Duranti, L. (2001). The impact of digital technology on archival science. Archival science, 1(1), 39-55.

[88] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436.

[89] Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural networks, 61, 85-117.

[90] Hinton, G. (2018). Deep learning—a technology with the potential to transform health care. JAMA, 320(11), 1101-1102.

[91] Hinton, G., & Lecun, Y. (2015). Guest editorial: deep learning. International Journal of Computer Vision, 113(1), 1-2.

[92] Payne, N., & Baron, J. R. (2017, December). Auto-categorization methods for digital archives. In Big Data (Big Data), 2017 IEEE International Conference on (pp. 2288-2298). IEEE.

[93] Loughin, T. M., Payne, N., Casilla, R., and Lum, C. (2017) "Statistical Modeling of Discrete Percentage Measurements With Application to Construction of Acceptance Bounds for Wood Failure in Structural Adhesive Testing," *Journal of Testing and Evaluation* 45(5

[94] Sarikaya, R., Hinton, G. E., & Deoras, A. (2014). Application of deep belief networks for natural language understanding. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 22(4), 778-784.

[95] Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. In Advances in neural information processing systems (pp. 649-657).

[96] Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759.

[97] Lai, S., Xu, L., Liu, K., & Zhao, J. (2015, January). Recurrent Convolutional Neural Networks for Text Classification. In AAAI(Vol. 333, pp. 2267-2273).

[98] Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., & Lindahl, E. (2015). GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX, 1, 19-25.

[99] Pal, S. K., & Wang, P. P. (2017). Genetic algorithms for pattern recognition. CRC press.

[100] Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2014). Data mining with big data. IEEE transactions on knowledge and data engineering, 26(1), 97-107

[101] Gerla, M., Lee, E. K., Pau, G., & Lee, U. (2014, March). Internet of vehicles: From intelligent grid to autonomous cars and vehicular clouds. In Internet of Things (WF-IoT), 2014 IEEE World Forum on (pp. 241-246). IEEE.

[102] Bonnefon, J. F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. Science, 352(6293), 1573-1576.

[103] Lemieux, V. L. (2016). Trusting records: is Blockchain technology the answer?. Records Management Journal, 26(2), 110-139.

[104] Lemieux, V. L., & Sporny, M. (2017, April). Preserving the Archival Bond in Distributed Ledgers: A Data Model and Syntax. In Proceedings of the 26th International Conference on World Wide Web Companion (pp. 1437-1443). International World Wide Web Conferences Steering Committee.

[105] Lemieux, V. L. (2017). Blockchain and Distributed Ledgers as Trusted Recordkeeping Systems. In Future Technologies Conference (FTC) (Vol. 2017).

[106] Lemieux, V. L. (2017, December). A typology of blockchain recordkeeping solutions and some reflections on their implications for the future of archival preservation. In Big Data (Big Data), 2017 IEEE International Conference on (pp. 2271-2278). IEEE.