Title

Designing Expert Systems for Archival Evaluation and Processing of Computer Mediated Communications: Frameworks and Methods

Author

Gilliland, AJ

Publication Date

2016

Peer reviewed

CHAPTER 25

Designing Expert Systems for Archival Evaluation and Processing of Computer Mediated Communications: Frameworks and Methods[1]

Anne J. Gilliland

Abstract: The third-party identification, evaluation, long-term preservation and retrieval of networked computer-mediated communications (CMC) such as electronic mail and social media have recently become subjects of much public debate. They also present persistent challenges for archivists. This chapter first offers a retrospective reflection on an applied research study that was conducted almost two decades ago investigating the possibilities of automating how university archivists appraise and acquire electronic mail. It describes the context of the study and the research design and methods that were employed. The latter included using bibliometrics to identify appraisal domain experts, acquiring and codifying knowledge from those experts, and the iterative development and testing of an expert appraisal system. The chapter then reflects upon what was learned from the study in terms of the utility of the methods and the aspects of this research approach that might remain useful for archival processing of documentation generated by social media such as Twitter, and email and cell phone communications today. It concludes by reflecting more broadly on how archival systems development research stands the test of time as technology evolves, institutional roles and conceptual frameworks shift, and methodological approaches gain or lose appeal.

Introduction

Which materials created through computer-mediated communications (CMC)[2] such as tweets, electronic mail, SMS (short message service), Facebook and blogs might be sufficiently valuable to posterity to preserve and make available in the future? What processes could be used to make that determination and acquire, manage, redact sensitive information, disseminate and retrieve those materials? Given the massive volumes and inter-relatedness, the potentially sensitive nature of some of these communications, and the usually very limited human resources of archives, how might aspects of those processes be effectively and appropriately handled automatically?

In January 2015, University of Oregon officials placed the head of special collections and the electronic records archivist on paid administrative leave following what the university claimed was the unlawful release by the archives of 22,000 digital documents that included confidential correspondence of the last four university presidents relating to faculty, staff and students. The University of Oregon is a public university and is subject to Oregon public records law, although the law does provide for certain types of sensitive information to remain confidential. The University Archives had acquired the documents as an electronic accession as part of its records management program. The documents were released upon the reference request of a University of Oregon economics professor before they had been individually reviewed and potentially redacted by the archivists for

any sensitive content. That professor subsequently uploaded one of the presidential memos to an online blog and the university began investigating how he obtained it.

By March 2015, the head of special collections was informed that his contract would not be renewed and the electronic records archivist had resigned her position. In April 2015 the former head of collections gave an interview to a local newspaper in which he claimed that the university had used him as a scapegoat. He maintained that "The library was being deluged with electronic archival material ... The mass of material was so great that there was no way special collections could vet or even organize it ... Archivists nationally are struggling with the same problem." A 2013 external review had in fact found that the library was indeed understaffed and the University Librarian had subsequently attempted to draw the university administration's attention to the insufficiency of the resources of the archives-administered records management program. The university countered the claim of scapegoating in a prepared statement: “Regardless of the Libraries’ infrastructure, however, it is the responsibility of the Head of Special Collections and University Archives to supervise the archives and records management unit, and to ensure that documents containing private and confidential information are properly reviewed and not improperly released.” [3]

The case provides a sobering opportunity to reflect upon the capacity of current records management and archival appraisal practices, archival processing and reference services, and automated processing tools to cope with the exigencies presented by high volumes of digitally created documents, especially when their content is not readily apparent. Indeed, it has led to a flurry of commentary among archivists on social media about where to turn for assistance in addressing the same conditions in their archives and asking where relevant research studies might be found. It also unfortunately underscores how, despite these issues having been flagged to archivists over two decades ago, they continue to be more pressing than ever. In that spirit, therefore, the first part of this chapter provides a retrospective overview of the historical context and methods of a study I conducted between 1993 and 1995 investigating the possibilities of automating how archivists appraise (i.e., arrive at decisions about whether records and other materials have archival value) and acquire electronic mail through the iterative development and testing of an expert system (i.e., a computer system that is designed to emulate the decision-making processes of a human expert). The second part of the chapter contemplates contemporary CMC contexts, values and needs, drawing on the example of the Twitter Archive at the Library of Congress, and discussing how the questions posed above continue to provide a rich vein for ongoing research investigation and not just in the context of university archives and records management. The chapter concludes with some more general thoughts on the role of systems development research[4] and how it stands the test of time as technology evolves, institutional priorities and conceptual frameworks shift, and methodological approaches gain or lose appeal.

Developing an Expert System to Appraise University Electronic Mail

Context of the study

In the United States, the critical need to conduct empirical research into the most effective ways to manage electronic records was recognized at the Working Meeting on Research Issues in Electronic Records, held in 1991 and sponsored by the U.S. National Historical Publications and Records Commission.[5] I began my study in 1993 (the year of the first public release of the World Wide Web), in the midst of some major conceptual shifts and research thrusts relating to recordkeeping and electronic recordkeeping more specifically. Alarms were already sounding about the need to assess the "recordness" and long-term values of the growing use of email and grapple with its preservation. For example, in the U.S., what came to be known as the PROFS lawsuit[6] regarding White House email and the U.S. National Archives' lack of records management oversight over it, had been wending its way through the U.S. courts since 1989 and would continue to do so for several more years; and at the University of Michigan, where I had been working as an archivist and doing my doctoral studies, another prominent lawsuit with archival implications was brought by a former student who sought to access faculty email under state freedom of information legislation. In Canada and the U.S., Terry Cook and David Bearman had recently separately published several "out-of-the-box" treatises on the management of electronic records that promoted evidentiary and systems thinking and ideas about post-custodiality.[7]

At the same time, however, academic research in archival studies was in its infancy,[8] with little in the way of international collaboration or even ready access to literatures from other countries. Trans-national cross-fertilization, as it influenced the American context, occurred more at the level of government archives practice, where archivists had occasions to interact with each other through various forums and initiatives, than in other archival settings or in academic programs. The American archival profession is large, diverse, and decentralized. The National Archives had not played the same central role in the framing, standardization, automation or integration of professional practices across government or other types of repositories in recent decades as had national archives in many other countries, and national archives staff tended to be more active in government and data archives associations than in the national and regional professional associations. There wasn't at the time, therefore, a large degree of interaction between government archivists, especially those who worked with electronic records and at the federal level, and the college and university (C&U) archivists, who comprised the largest sector of the American archival profession. Moreover, C&U archives had a strong tradition of autonomy from each other and from government archives.

My epistemological stance at the time was rooted in a belief that the implementation of descriptive standards, automated processes and archivally-oriented information retrieval could "normalize" what were often poorly articulated and idiosyncratic archival practices. I saw the benefits of such an approach being that it might nudge the U.S. archival profession towards the new behaviours and explicit conceptualizations necessary to address the challenges of born-digital materials and to exploit the networking and computational capabilities of information technology.[9] This stance was undoubtedly influenced by the optimism prevalent at the time that information technology and standardization could make sweeping and rapid changes in ideas, institutions and practices that might have centuries of social and cultural embeddedness. In recent years,

however, the limitations of this stance have become increasingly apparent to me and I have incorporated a more anthropological and bottom-up approach toward the a priori analysis of records creation and recordkeeping systems and metadata. I use ethnographic and community analysis methods to try to understand in socio-cultural and political terms why these processes, systems and metadata came about, how they operated, and to whose benefit and whose disadvantage in order to generate a lens for determining feasible and context-appropriate parameters for systems development.[10] This latter inclination in part informs how I have chosen to put this chapter together. Nevertheless, my grounding in systems analysis and information retrieval based in information science and business information systems taught me the value of rigorously defining concepts that needed to be operationalized in context when working with both human and digital systems, and I find myself constantly returning to that as a reference point in my research.

My professional and research engagement in the early 1990s had both been directed toward addressing how to identify ways to assess the potential archival values of materials created by CMC and for C&U archives to acquire any deemed to have such value. In 1993 I had just finished working on the Bentley Historical Library Computer Conferencing Appraisal Project, a federally-funded research project at the University of Michigan that developed approaches for appraising, accessioning and automatically describing active and inactive "computer conferences," an early form of social media.[11] That project was strongly influenced by Helen Samuels' ideas about documenting academic environments.[12] I wished to build upon that experience by examining how a university archives, with limited technological and human resources, and a consciously less clearly defined recordkeeping mandate than most government archives would have, could automatically appraise and acquire electronic mail generated in the course of university activities. At the same time, I was frustrated by the number of assertions that I encountered in expositions of archival theories and practices in the professional literature that had never been tested to ascertain their validity or efficacy over time and in different contexts. This frustration drew me toward a method--expert systems development--that required precise conceptual articulations of those assertions and that provided both a feedback mechanism and a real-time accounting of its own efficacy. I was not alone in this impulse for precision and clarity. The Pittsburgh Project, led by Richard Cox and David Bearman, was underway at the same time. It generated a number of "production rules" that were designed to be built into electronic recordkeeping systems to ensure the creation of trustworthy, segregable and preservable records.[13] Similarly, the Preservation of the Integrity of Electronic Records Project (a.k.a. the UBC Project) led by Luciana Duranti, Terry Eastwood and Heather MacNeil, used IDEF0 modeling to depict graphically and unambiguously the workflow, inputs, outputs, constraints and resources involved in creating reliable records in electronic systems.[14] However, in both cases (which constituted some of the earliest North American collaborative research led by archival studies academics) what were supposed to be clear articulations of archival and recordkeeping ideas still proved to be too arcane for systems designers and even archivists to follow with ease (see the chapter by Hofman in this volume for more on this issue).

The first ARPANET network mail message was transmitted by Ray Tomlinson between

two machines side by side in a lab in 1971. By 1993, CMC technologies had been available for over twenty years and electronic mail was by far the most prolific and prominent application. However, even though there was considerable ferment in the field over other forms of electronic records and recordkeeping, the challenges and potential of the materials created through the use of CMC had only obliquely been addressed by both archival practice and the professional literature. Various government agencies published draft guidelines on the retention of electronic mail[15] but actual implementation of these guidelines remained a problem. This was largely because of the legal, technical, and political difficulties that these materials presented to archivists. Examples of these difficulties included defining what might legally be considered a record in a given environment; how best and when to capture and appraise such transient materials; and how to avoid violating, or being perceived by email users to violate, their personal privacy.

Until litigation started to occur in the United States regarding the possible record status of email, there had also been a pervading sense among administrative users and archivists alike that such electronic communications were employed only by a limited sector of organizations and not used for the important administrative activities traditionally falling under the purview of records management or documented by existing archival programs. Moreover, many administrators and archivists viewed email as a convenient informal and informational way to communicate, similar to the use of the telephone. As they do today, individual emails frequently contained a mixture of official and personal business that in itself posed major problems for archival appraisal, arrangement, and description.[16] Since in most cases individual emails were created within a system that did not make distinctions between the kind of business or personal function in the context of which the email was sent, it was felt that appraisal, if it was going to take place at all, would need to be conducted not of the system as a whole, but in some way at the level of the individual materials that it transmitted and stored.

At the same time, because of how it had enhanced scholarly communication in the academic setting, computer-mediated communication and its impact on the research community and society in general, had become a subject of study in itself.[17] The influence of documentation strategists within the archival community was also beginning to be felt in the world of electronic records management. Advocates of documentation strategy approaches[18] were quick to point out that archival involvement in systems design and analysis could bring with it increased opportunities to document not only official activities, but also issues and movements of social and topical value. Terry Cook had noted in 1991 that: “Policy files ... suffer from underdocumentation because important decisions are made by telephone, personal conversation, or in other ways.”[19] Electronic mail was, increasingly, one of those “other ways.” Helen Samuels argued that the spectrum of college and university activities had been poorly documented by archivists.[20] She suggested that several core functions, including intellectual life and socialization, were poorly represented in the administrative records traditionally acquired by C&U archives and in 1992 published *Varsity Letters*, a blueprint for documenting academic environments.[21] The findings of the Bentley Historical Library Conferencing Appraisal Project indicated that these functions were more completely documented within computer

conferences, supporting speculation that this would likely also be the case with email.

A major and immediate problem for all archives, and especially for C&U archives remained how to undertake such an appraisal task, especially with a small staff and often with little or no available technical expertise. Electronic mail not designated for archival retention at, or soon after the point of its creation was unlikely to endure long enough to be appraised within its original context because of system purges and purposeful or accidental deletions by users. However, as the Bentley Project had also found, the problem was complicated by the concern that if archivists were able to identify a mechanism for designating email for archival retention at the point of original transmission, they might somehow violate the privacy of correspondents who legally and ethically would have expectations that their current email would be both confidential and secure from outside viewers except with their explicit knowledge and consent.[22]

This exploratory study, therefore, represented an empirical effort to identify an inexpensive, readily comprehensible automated way to identify documentation of long-term value from the ever-growing mass of electronic mail created, communicated and received within a major public university.

Research Design

In line with the research in communications and organizational behaviour on which I was drawing at the time, the underlying philosophy of the research design was to blend qualitative and quantitative techniques in such a way as to strengthen the reliability and generalizability of the resulting research outcomes. [23] Although it built on prior work with the Bentley Conferencing Appraisal Project, I did not model the study directly from any one approach used previously in archival research. I considered and ultimately rejected a number of possible conceptual and systems approaches, chief among them was that being taken by the Pittsburgh Project, which was based around the transactional nature of business communications.[24] While this transactional approach was derived from standard business practices and regulatory needs, it did not completely correspond to how the law might interpret a record in an electronic mail environment. Instead its value lay in the fact that it afforded archivists and systems developers a way to identify a record and the opportunity to develop archival functional requirements for electronic mail systems that operated at the user interface level. At this level, Bearman argued, users would be empowered to identify their outgoing and incoming mail according to a menu of different genres and document usages, some of which would be scheduled for archival retention based on their form-of-material.[25] In order to capture email in the electronic environment, therefore, this approach offered two options: establish an institutional information policy that mandated appropriate disposition for specific types of record material being transmitted; and/or build a mechanism into the architecture of the email system for email users to identify their own transactions according to their administrative value (e.g., filing and routing menus), as has been done at the World Bank and the Northern Territory Department of Mines in Australia. The major limitation of this approach in terms of its relevance for this study was that it was directed toward the capture of official materials, the scope of which can be harder to ascertain in institutions such as universities where

records tend to be less defined by legal and regulatory mandates, and where materials created by faculty members and sometimes even prominent administrators are often treated more like personal papers than official records. Moreover, Samuels' research had exhorted C&U archives to assess a much broader base of university documentation than just administrative records. The Pittsburgh approach would likely miss less formal correspondence, as well as sequences of email interactions between pairs and clusters of individuals. It also left opportunities for email users to misfile documents accidentally or deliberately, possibly leading to incorrect retention and destruction actions.

I decided instead to develop an automatic appraisal mechanism in the form of a prototype expert system that could function as a "front-end" to a specific university email system, but that was potentially extensible to any stored file of undifferentiated CMC, in order to assist institutional archivists in the appraisal process. Such a front-end system would facilitate the appraisal of both official and personal communications. At the same time, it would distance not only the email user, but also the human archivist from the active record and thereby mitigate some of the threats to personal privacy perceived by individual email users, and any risk that archival "monitoring" of electronic communications might affect the nature of the interaction by making the participants self-conscious or feel intimidated.[26]

This approach was suggested by a project that developed in the mid-1980s as an engineering and business application at the Massachusetts Institute of Technology (MIT). MIT's Information Lens Project was a rule-based information management system that emphasized the development of customizable filters or templates that could be used by individual or group end-users of electronic mail to handle their communications and information overload more effectively. These filters were built around what they termed "cognitive," "economic," and "social" factors that I suspected had many analogies to the criteria and values used by archivists during the appraisal process. The expert assistant I envisaged would contain a series of progressively refined and customizable filters and profiles designed to deselect both individual email messages and specific types of email of no archival value. The filters would be based upon a set of rules codifying expert views on appraisal, but defined in the context of the mandate of the individual archives. They would act upon analyses of the form, body (i.e., content), and header, routing and signature file information contained in the messages, apparent message trails, and knowledge about the positions and status of email senders and recipients drawn from the university's standardized electronic directory system.[27]

The study assumed that, as with manual appraisal, a high percentage of active email would probably be filtered out automatically as having little or no archival value (e.g., listserv messages, routine memoranda), and the remainder that was identified as potentially having long-term value could be digitally stored for three or more years. After that period, archivists could subject the stored material to a more thorough manual appraisal, which would assure, with the benefit of hindsight, the retention and description of only those materials deemed to exhibit values in line with the university archives' mandate and appraisal policies.[28] (Of course it should be noted that today, as the University of Oregon case highlights, even if over 90% or more of email were to be

filtered out, the remaining percentage would likely still be too voluminous for manual processing and further computational approaches such as those being attempted with the Twitter Archive and discussed later in this chapter would need to be applied.)

In order to carry out the research, a bibliometric analysis of citations in the archival literature (a quantitative technique) was first conducted in order to identify living individuals who might be considered "experts" in the area of archival appraisal. The choice to use bibliometrics was in part because I had already used the method in other research and therefore was familiar with what it entailed and also its limitations. It was in part also a political choice. I wanted an “objective” way to identify figures whose ideas about appraisal had been influential upon the archival field in North America that would withstand challenges as to why others had not been asked to participate in this study. And I wanted to work with these figures in particular because I was also interested in whether a codification of the range of prominent appraisal ideas was even possible.

Since the North American archival literature was relatively small and in-bred,[29] I sought to augment the identification of experts from the citation analysis using a snowball sample to ensure that all possible experts would be identified and contacted. In line with accepted thinking regarding knowledge acquisition from experts, I conducted focused written and oral interviews and follow-up discussions with the identified experts (a qualitative technique) to ensure that theoretical expert knowledge was adequately represented in the core and alternative rules included in the expert system rule-base. Evaluation of the various development stages of the prototype expert system were based upon archival practitioners’ expert appraisal experience and knowledge (that is, the appropriate university archivists) in the context of the specific mission, needs, and culture of the institution whose communications were being appraised (qualitative), as well as by examination of the statistical data regarding the ratios of filtered and non-filtered electronic communications (quantitative).

A modular approach such as this was valuable when conducting an exploratory study in areas about which little knowledge or data existed, and where consequently there was a danger of developing a study that could have both low validity and low generalizability. Each of the three research phases (that is, the citation analysis, the knowledge acquisition process, and the building of the expert prototype) had considerable research value in itself and was designed to stand independently of the others. This modular approach allowed possible confounding variables to be identified in the research phase where they first appeared, and also ensured that the research was not a wasted effort should subsequent modules fail to produce successful results.

Research Methods and Related Procedures

a. Bibliometric citation analysis

Bibliometrics is a quantitative research method that belongs to the same family of “metrics” methods as sociometrics. It is used, through mathematical and statistical analysis of citations and sometimes of content, to discern patterns and draw inferences

about the influence and dispersion of particular authors, publications and subjects within a given literature or literatures, and by extension, field or fields. Since its inception, certain bibliometric “norms” or “laws” have been discerned, largely based on scientific and technical literature analyses, for how literatures within fields are expected to behave. Bibliometrics does not, however, address biases inherent in citation practices, and the data analysed is not always complete or accurate. Logistically, its biggest limitation until recent developments in automated citation identification and indexing on the web, was that citations had to be identified and then analysed by hand (as was the case for this study).[30] This was both time-consuming and did not scale well when large amounts of documentation needed to be analysed.[31]

In bibliometrics, researchers must unambiguously define and then rigidly follow their own rules about how a given literature is selected and analysed. Declaring these rules in writing up one's results is also important since it enables others to replicate the study or the study's approach using with subsequent or similar citation sets, thus allowing for cumulative knowledge development about a field. Since this study was concerned with codifying archival appraisal as it was understood and practiced in North America, only the North American monographic and periodical literature was analysed, and 1972 was selected as the likely first date when authors on appraisal might have begun to be aware of CMC. [32] There was no benchmark in archival research for developing citation analysis guidelines other than that the guidelines be explicitly stated and consistently followed so that the analysis could be replicated by other researchers.[33] Rules for analysis, therefore, had to be established for this study.[34] [35]

b. Expert knowledge acquisition

Living authors whose works fell in the top 25% of citations identified were considered to be experts in the field of archival appraisal and were contacted and asked to participate in this study.[36] These individuals were sent a letter explaining the nature of the study, asking for their cooperation, and giving a short list of broad questions and issues regarding archival appraisal for their consideration. The questions asked about what the experts regarded as the central tenets of archival appraisal theory, which texts they believed to be particularly valuable, how they personally might approach the appraisal of electronic records, what practical factors should be considered together with archival theory during the appraisal process, including any considerations that they felt might be specific to the college and university environment, and the names of those individuals whom they regarded as experts in the area of archival appraisal. The process of knowledge acquisition involved a telephone or in-person interview (which was recorded in cases where it was both feasible and agreed to by the experts) or written responses structured around the broad questions and issues introduced in advance by the letter to the experts. This knowledge acquisition process was only semi-structured and used open-ended questions in order to encourage candid comments and free-flowing thoughts on the part of the experts as to the rules, guiding principles, and heuristics they used in archival appraisal. It also left me free to follow up on any points that were unclear or insufficiently articulated to be codifiable.

Any individuals identified by these experts but not by the citation analysis as being influential in the area of archival appraisal (perhaps because newly published works had not yet been significantly cited, or because an individual did not publish but held a key position as an appraisal archivist), were also considered to be experts and were contacted with the same letter and questions, and the same interview process was conducted. The knowledge acquired from each expert was codified into a set of unambiguous statements and was sent back to the expert for his or her review, clarification and comment, and then revised again. Sometimes this process was repeated several times. This ensured that the knowledge accurately represented the views expressed by the experts as well as giving them an opportunity to add further thoughts they might have on the subject.

c. Building and testing the expert system

The decision to develop an expert appraisal assistant was based on several characteristics exhibited by expert systems that made this technology particularly appealing for a project of this nature. Expert systems were widely employed (and remain so today) in situations where complex reasoning and expert knowledge are involved, for example, in business decision-making, marketing, and scientific and medical diagnostics. In particular, I hoped that the ability of expert systems to make decisions when faced with uncertainty should allow archivists to build a system that not only included any categorical principles that guide archival appraisal theory, but also the kinds of institutional and even personal heuristics that inform archival appraisal practice (although the reasoning abilities and potential to function as useful assistants of expert systems can only be as good as the extracted and coded knowledge of which they are built). There were several additional advantages to taking an expert systems approach that were not necessarily the case in other forms of systems development. These advantages included the facts that expert systems could:

- be developed and used by individuals who are not highly skilled in computer programming (in this case, even by archivists themselves);
- be cost-beneficial in that they require relatively little investment in terms of time, staff, and software, with the potential of a large pay-off in functionality;
- be rapidly prototyped and easily modified, especially small to mid-sized systems that utilize up to approximately 200 rules;
- be constructed using a variety of off-the-shelf expert "shell" [37] software rather than original programming, if desired, or a mixture of both, allowing for a high degree of customization;
- provide a readily customizable interface in addition to a knowledge base and an inference engine, which can make a system very accessible to archivists who do not have much technical background;
- provide various feedback mechanisms that could elucidate the underlying appraisal decision criteria and processes; and,
- interface as "front ends" to other systems such as databases or telecommunications systems.

The major limitation of expert systems is that they are “brittle” in that they do not know what they do not know. They need to work, therefore, within carefully drawn subject or functional limits. Without such limits, they tend to crash ungracefully when faced with a situation for which they were not programmed. As a result, expert systems technology would not be appropriate for use in a situation where a system would be required to operate alone twenty-four hours per day and where a poor, or simply wrong decision might cause a major accident or incident. In the case of archival appraisal, however, I decided that the needs to have the system running constantly and to make a “correct” decision regarding every single email were not crucial and therefore that the above limitations did not pose sufficient reason not to use expert technology (perhaps in light of the University of Oregon case I was mistaken about that!).

I selected an expert system shell based on the opinions of several individuals knowledgeable in the area of knowledge engineering, reviews of existing expert systems, and necessary technical specifications. I also believed that it was important that the software have explanation facilities, such as understandable rule-tracing, in order that the reasoning behind expert appraisal decisions could be deduced and explained to a court of law, other archivists, or concerned research communities. At the same time, however, one of the points of the study was to look at the extent to which an archivist could articulate archival requirements of an expert appraisal assistant to a systems designer who would then build it, without needing to become intimately involved in the details of coding the system.[38]

Aluri and Riggs had outlined several stages involved in building a library expert system (and these are largely in line also with those outlined by Nunamaker and Chen as typical in systems development approaches:1) identification of a topic or area; 2) conceptualization (for example, what will be the system's parameters, who will be the experts and engineers, how will the work flow be organized?); 3) formalization (decisions regarding the tools, concepts, and design to be used); 4) implementation of a prototype; and 5) evaluation of the system.[39] These stages, which are fairly generic for the development of an expert system in any environment, were largely followed in this study. Today there are software applications that will interview experts and automatically code the knowledge derived thereby. However, this process still needed to be conducted manually at the time when this study was being undertaken

As already indicated, the prototype developed for this study served as an expert assistant in the appraisal (more properly, an expert assistant in the “meta-selection”, because it was making a preliminary selection and not a final appraisal) of electronic mail in a public university setting. Although it could potentially function proactively as an archival front end to an institutional mail system, or its rule-base could be further customized to serve administrative or personal information management purposes, it was never intended to function completely independently of human review. This prototype was customized to address the mandate and appraisal practices of one particular university archives in a major American public research university.

Methodologies advocated in expert systems literature for testing the effectiveness of prototype and operational expert systems were very scant at the time and lacking in rigor. On the whole they were not relevant for the nature and use of this prototype system.[40] Instead, I devised my own evaluation strategy, employing the assistance of the university archivists who were responsible for traditional appraisal of official institutional correspondence to assess a sample of messages being caught in the filters, and then refining the filters accordingly and retesting them. In line with the iterative development, evaluation and refinement processes that underlie systems development as a method, this strategy needed to be modified several times as the project progressed. Nevertheless, the university archivists highlighted several interesting difficulties in making manual judgments due to the volume of messages, the difficulties of making accurate judgments at the level of individual item and with what they considered to be technical content. Moreover, because the messages dated from 1989 to 1994 and had not been redacted to reduce their potentially sensitive nature (i.e., I did not have permission myself to view them individually, as discussed further below), I chose to examine the results of the filter combinations only in terms of changes in the ratios of items marked for retention and deletion with different filter combinations.

Reflections on the Methods, Extracted Knowledge, and Test collection Used

The citation analysis proved to be an excellent mechanism for providing some insight into the nature of the archival field at the time. It indicated that the North American archival profession had a very small, and rather self-referential corpus of opinion leaders in the area of appraisal, and that there was very little influence from literatures in other fields (including technological fields) or archival traditions. Instead, many authors published only a few thoughtful and thus heavily cited articles or monographs in their career but did not pursue a consistent and rigorous path of research investigation. This would likely be different today, since there are now international research journals with broader coverage, and an identifiable cohort of researchers in North America and around the world.

Eight of the experts identified agreed to participate in the study. Together they represented a range of ages, backgrounds, and proponents of differing appraisal strategies, something that initially I considered to be optimal. A larger number would have made the codification process considerably more difficult. However, most expert systems only attempt to represent the knowledge acquired from one individual and as it transpired, and as guidance on expert knowledge acquisition cautions, heterogeneity among the experts due to being drawn from different archival traditions (i.e., U.S. and Canadian - see chapter by Gilliland, Lian and McKemmish for further discussion of this consideration), different theoretical approaches (e.g., macroappraisal, Schellenbergian) and different institutional backgrounds (e.g., university archives and special collections, government archives, historical society) proved to be a significant limitation in defining a definitive rule base. With hindsight, another way in which this process could have been handled, if the experts had been available to do so or if the system had been available to them remotely, would have been for them to interact with the rule base directly and make corrections “on-the-fly. However, that option was not available at the time. Alternatively,

a Delphi process could have been used. Delphi studies are used to reach consensus or predict an outcome or correct answer. They most commonly involve asking experts to respond to a series of rounds of questionnaires. After each round, the facilitator summarizes and anonymises the experts' responses and the reasons they gave for those responses and sends them out to the experts, encouraging them to revise their prior responses in light of the summaries and reasoning provided. The benefit of such an approach for developing an expert rule-base would be to obtain a single consensus about rules. While this might have improved the technical outcomes of the study, however, it would not have elucidated the range of opinions on appraisal--something that was a topic of interest for this study. Moreover, similarly to any standard that is developed through a group consensus development process, it would likely be less responsive to individual institutional or local contexts, and this study was trying to find a balance between broad agreements and local mandates and practices regarding appraisal goals, values and processes.

Based on the knowledge acquisition process that took place through interviews and written correspondence, a wide range of thoughtful comments on appraisal emerged that demonstrated the complexity and sophistication of appraisal. This knowledge acquisition process also demonstrated a considerable lack of consensus between experts on why and how appraisal is conducted that was strongly based on their levels of experience and milieus. In fact, when I grouped the acquired knowledge into 45 principles and heuristics, only 2 of these were expressed by all the experts. Most of the experts who participated in this study, as is probably the case with archivists in general, were extremely comfortable expressing their ideas using text and drawing upon the richness of language to convey the subtleties of the art, as well as the science, of appraisal. Complex and sophisticated knowledge expressed in this way can be a double-edged sword, however, when it comes to developing a system to emulate human processes such as appraisal. While such knowledge can lead to the development of a very powerful system, a lack of clarity, precision, or consistency in how it is expressed to a designer, as well as inherent system limitations, can result in a system that has to resort to using simpler, less controversial concepts. Today another way to go might be to opt to examine the heuristics of personal file management, beginning with studies that have been published in the growing literature on personal digital archives. When one builds one’s own email filters and filing schemes, in some ways that is analogous to building one’s own expert system to appraise and classify one’s email.

After finding so little commonality among the appraisal experts, I could have ended my research with those findings, but I did not believe that such a finding negated the needs originally outlined for developing automated appraisal front ends for electronic communications, nor that it made it impossible to develop an expert appraisal system tailored to one specific institutional archives. Because of the lack of agreement on appraisal principles and heuristics (i.e., rules of thumb), I modified and considerably simplified the original design and combined the acquired knowledge into several groupings representing the type of appraisal considerations cited, as well as how frequently each consideration was cited.

Unlike the Bentley Conferencing Appraisal Project, which worked with both live and "archived" conferences, the prototype was iteratively tested using an extensive test collection of electronic mail of academic and research provenance associated with one senior administrator within that university. The email had been downloaded and a certain amount of redaction had taken place to address sensitivity and privacy concerns. I had sought out as extensive, heterogeneous, and organizationally cross-sectional source of academic administrative and/or research electronic communications as possible. For political and legal reasons, obtaining a test collection for this study, whether live, or offline and "massaged," proved to be possibly the most problematic aspect of the entire project. While far from ideal for the actual systems development and testing, I was very fortunate to have been given the opportunity to work with this particular test collection. This issue, however, points up two very important issues relating to archival research infrastructure as well as generalizability and validity concerns for research outcomes. The first of these is the essential need for test collections with which those engaged in archival systems design can work. The second are the benefits and disadvantages of artificial and real-life test collections. The information retrieval (IR) community has a long history of working with artificial test collections (i.e., collections of materials or bibliographic data specifically constructed for conducting retrieval tests and where every item is known). These allow for accurate assessment of such aspects as recall and precision, replication of empirical studies, and subjecting the same test collection to a barrage of different research approaches (see Furner and Gilliland chapter in this volume for more discussion of IR research). However, such collections can be far removed from the complexities and idiosyncrasies of a real-life system and its contents, where also the content can be surmised but often not accurately delineated. While using a real-life system and its contents can make for more realistic studies, it can be difficult to evaluate the efficacy of the tests being carried out, especially in terms of measurements because its contents, parameters and characteristics might not be completely determinable. It might also be hard to identify variables that could be affecting evaluation results. The test collection to which I had access unfortunately combined the limitations of both artificial and real-life test collections.

A further note on the limitations of my test collection is also warranted here since recognition of these limitations in the context of this study is, in itself, a valuable component of the research. The most obvious limitations in this study (and also considering Samuels' documentary objectives that inspired it) was that the test collection only contained the outgoing messages, and therefore reflected the electronic mail creation patterns of only one key individual within the university. A major characteristic of email, and many forms of materials created by CMC, is the way in which they messages are connected as trails of correspondence, or by clusters of senders and recipients. Any automated archival system needs to be able to comprehend and exploit these relationships because they are an important aspect of the documentary context of those materials. For this reason, testing on a live system would be preferable to that of a limited and massaged test set. Moreover, when the messages in the test collection had been saved as ASCII text, they were saved as they had been seen through the email viewer and were stripped of much of their routing information, again essential evidential detail for assessing records (it should be noted that this would also be a matter of concern for real life acquisitions if

the creators had massaged or redacted them significantly before transferring them to the archives). The research data security protocol that I was required to follow also provided an additional challenge that it is important should not occur in real-life developments of this sort (where the archivist or systems developer would presumably be employed by the institution) in that it necessitated that the researcher and programmer design the expert system without looking at the format or contents of the messages in the test collection.

Possible Roles for Automated Systems Development in Contemporary Third-party Identification, Evaluation, Long-term Preservation and Retrieval of CMC

Although in the interim much has changed in the archival field and with CMC technologies, communications infrastructures, and the ubiquity and nature of their usage (see Acker chapter in this volume), this study remains one of a very small corpus applying systems development as a research method to address the archival problems associated with long-term management of and access to CMC and indeed to understand the nature of CMC and the associated behaviours and patterns of their users/creators.[41] This is particularly surprising given the fact that traditional correspondence, and now digital correlates such as electronic mail, text messages and tweets are often viewed as rich sources of both information and evidence by researchers and indeed investigators of all kinds. Moreover, the processes that are or might be applied in the evaluation, capture, and potential secondary uses of CMC by parties other than the creators or authors of those media have been the subject of several public controversies in recent years. Prominent examples include the release by Wikileaks in 2010 of more than a quarter of a million classified State department cables (known as Cablegate); and revelations made by Edward Snowden since June 2013 that the U.S. National Security Agency (NSA), as part of its terrorist surveillance program, was provided unsupervised access to all fiber-optic communications, including electronic mail and text messaging conducted using major American telecommunications providers, and associated metadata.[42] Such cases have not only raised questions about whistleblowing and unsanctioned release of privileged or sensitive communications, but also about how any organization, including Wikileaks and U.S. intelligence agencies, might be in a position manually or computationally to gain access, review and, when necessary, redact, such a high volume of materials before release, as well as how it might arrive at assessments about what to release and what to redact.

Since 2010, Wikileaks' activities have escalated and so has the volume of materials being released. In July 2012, more than five million leaked electronic mail messages from Stratfor, a geopolitical intelligence firm that provides strategic analysis and forecasting to individuals and organizations around the world, were released. In the same month Wikileaks also released over two million electronic mail messages between Syrian political figures, agencies and companies. Undoubtedly, when faced with such a high volume of materials that individually or in relation to one another may reveal unanticipated information and evidence through their content or communication patterns, there is a role to be played by automated systems that can assist with the appraisal or evaluation and possibly even the redaction of the documentation. However, for those interested in uncovering information and evidence from such documentation, for

example, the news media or intelligence operations, there is also a role for developing systems that are able to do more sophisticated forms of evidentiary retrieval.
While these examples have occurred for the most part outside the archival purview (although they have attracted much archival commentary), and presumably also open up venues for the kind of research discussed here in other than strictly archival applications, they also provoke many questions about the archival evaluation and processing of records generated by CMC and indeed what constitutes an archive and what records in the context of CMC. Some of these are illustrated by the example of the Twitter Archive at the Library of Congress, which I will discuss in some detail in this section.

Foreseeing the rise to prominence of CMC, Canadian archivist Catherine Bailey wrote in 1989 that:

> Electronic mail is the nearest written equivalent to the correspondence of the pre-World War II era, when decision makers committed their thoughts, feelings and judgments to discursive prose in official letters. The letters conveyed information for an immediate purpose; they were not written with an eye to history, nor did they serve the purpose of most official letters today--after-the-fact confirmation of decisions already reached.[43]

In 2013, the Library of Congress' justification for its decision to preserve or "archive" the Twitter Archive resonated with the same sentiments:

> As society turns to social media as a primary method of communication and creative expression, social media is supplementing and in some cases supplanting letters, journals, serial publications and other sources routinely collected by research libraries.
>
> Archiving and preserving outlets such as Twitter will enable future researchers access to a fuller picture of today's cultural norms, dialogue, trends and events to inform scholarship, the legislative process, new works of authorship, education and other purposes.[44]

Twitter is one of the most prominent examples of how, and how fast, ever-evolving forms of computer-mediated communications have transformed not only the ways and scale at which humans communicate with each other, but also the milieus in which they do so and the influence they can exert. The future of CMC and electronic mail appears today to be bound up with social media platforms and mobile operating systems like Android and Apple that support messaging (combining instant messaging, texting and electronic mail) and engaging multiple stakeholders, new pricing and market structures, protocols and standards. While there is debate on just how valuable the Twitter Archive will turn out to be, and also about the wisdom of acquiring social media through such a firehose approach and with little clarity about how, if ever, widespread access might be supported, there is no doubt that embedded in other forms of CMC, such as email, interactive web pages, text messaging, Skype sessions and blogs, is evidence in the form of the traces of and reflections on contemporary decision-making, reporting, conversations and other

activities. Previously such traces and reflections would have been contained in the formal and informal correspondence, reports, diaries and notes that today fill the shelves of archives and provide some of our richest insights into these activities. This is not to say that direct equivalencies can be drawn between the communications media of today and those previously in terms of how they are used and by whom, but rather to argue that much of what we are likely to value most as future historical documentation of organizational and personal activities and decision-making is now distributed across a web of interdependent and otherwise linked digital "documents."

In keeping with its mission to "acquire, preserve and provide access to a universal collection of knowledge and the record of America's creativity for Congress and the American people,"[45] in 2010 the Library of Congress in Washington, D.C. entered into a controversial partnership with Twitter and social data provider Gnip to build and preserve an archive of tweets. This corporate partnership arguably suggests the dawn of a new kind of archival access model. Gnip is currently the only third party developer that has access to all the firehose data from Twitter. Other vendors, clients and the growing corpus of researchers investigating social media patterns and implications must pay for access to tweet metadata through their tools and algorithms, or ping the public application programming interface (API) and hope for the best, a situation where access costs and availability of accumulated data not only limit such research, but also the replicability of its results and the testing of findings over time.

The Twitter Archive has received considerable media attention in terms of the strategies that are being employed by the Library of Congress and its partners, as well as the amount of public money that has been used to preserve the content of a form of CMC that is seen to be primarily trivial or over-hyped in terms of its influence by some, and to be a revolutionary tool capable of supporting private speech, citizen journalism, community organizing, and tracking world events such as elections, military conflicts, and pandemics by others.[46] Either way, it may be several years before its impact today can really begin to be understood. Its high profile and the associated debates about value have added to the existing pressure on American archives to justify or re-think the continued relevance and effectiveness of their traditional role and practices, and the values around which their judgments are based. It challenges archives to respond more proactively to the ways in which documentation is being created and might be used in the digital, networked world, lest they cede their role as preservers and providers of essential recorded evidence to others who are more technologically innovative. For example, would a national archival repository, rather than the world's largest bibliographic institution, similarly have determined that public tweets were worthy of acquiring and if so, on what basis, when and according to what processes? How well would their practices hold up to the kind of public scrutiny that the Library of Congress has experienced? Does the Twitter Archives' approach suggest that archival appraisal,[47] and perhaps even archival description are no longer either relevant or, indeed feasible? However, if appraisal is indeed an obsolete approach, then why has the Library of Congress found that it needs to justify its commitment to preserve and make the Twitter Archive available in terms of future research value and the costs entailed and what evidence can it draw upon to support such justifications?

As of January 2013, the Twitter Archive had completed retrospective accessioning of public tweets from Twitter's founding in 2006 to 2010, and was digitally ingesting[48] the tweets that had been sent each day subsequent to 2010 (almost half a billion a day and growing in January 2013).[49] The Twitter feed is acquired in real-time (i.e., without any lag time between tweeting and ingestion), and without going through any appraisal process (i.e., no selection mechanism is used, for example, to identify only tweets associated in some way with America). The Library of Congress argues that, "It is clear that technology to allow for scholarship access to large data sets is lagging behind technology for creating and distributing such data. Even the private sector has not yet implemented cost-effective commercial solutions because of the complexity and resource requirements of such a task."[50] In place of archival description, as already discussed, data mining of the chronologically-organized Twitter Archive is being conducted by Gnip to identify ways in which the current content of over 170 billion tweets and their associated metadata might be automatically discovered and retrieved. The Twitter Archive is also being made available to other commercial parties such as financial services providers, marketing companies, and social monitoring and analytics firms to mine, use for predictive modelling, and extend with additional services and capabilities. As yet, however, there is no public access to the Twitter Archive.

The massive volume, digital format, brevity and high variety (e.g., many tweets have embedded links or pictures) of individual tweets, as well as the high degree of inter-dependency and relatedness that often exists between them (i.e., their documentary context), lend a lot of heft to arguments in favor of abandoning traditional library selection and cataloguing processes, and relying on computational power to perform aspects of those functions as necessary.[51] One wonders whether, if there were more investment by and expertise within archives to develop their own access tools and algorithms, such commercial fee-based approaches and the limitations to use that can result might be averted.

Even though the Library of Congress sees itself as addressing a set of data rather than evidence management concerns,[52] keeping everything and relying upon computational processing to manage archived tweets and make them discoverable or compilable is a strategy that supports the capture and (re)presentation of a record (in the sense used by the Library of Congress) that is not static or isolatable but is, rather, continuously accumulating and evolving. It is, in fact, an approach that was advocated as long ago as 1991 by David Bearman, archives and museum informaticist, as well as some prominent information scientists. It was not taken up by archivists at the time,[53] but today is very much in keeping with continuum ideas about the dynamic and interactive nature of recordkeeping and use.[54] In line with methods used in social network analysis,[55] this computational approach also offers new capabilities to end users that were not possible when working in a physical paradigm, such as following threads, trails and patterns of tweeters with particular personal profiles, tweets and retweets; identifying dispersion rates for events and other phenomena discussed in tweets; and potentially visually mapping these by date, geography or influential tweeters, or correlating them with other digital traces of particular events.[56]

Some Concluding Thoughts on Applying Systems Development Methods in Archival Research

Since I undertook this research there have been several major shifts in archival thinking, and I have moved out of practice and into academia. Jerry McDonough has commented that academics have space and time to think and reflect. They can also look in places that practitioners cannot.[57] When I was doing this study, reflexive approaches[58] to systems analysis and development were not a major consideration and Donald Schön's work was just beginning to become influential in library and information science. However, even though systems design requires definitional precision, it still hardcodes the biases of the developer and the development context into the system, and every researcher in the field should attempt to acknowledge and account for that (see Evans' chapter in this volume). Ideally, conceptual work should come before implementation, but in a fast-paced technological world, there isn't always time for that conceptualization to occur or to mature, which is one reason why more of it tends to occur in academia than in practice. It should be noted also that methods and frameworks themselves go in and out of fashion. When I started this study, expert knowledge extraction was popular and American archivists still believed that they could use appraisal and life cycle management to cope with the growing volumes of electronic materials. Although they evolve more slowly than the current rate of technological development, this study illustrates that archival and recordkeeping frameworks, ideas and methods, and even institutional roles and mandates will also not last forever, but they do have an essential currency in the present, and systems development such as that discussed here is rooted in that present.

Systems development also has a lot of moving parts and someone invested in this kind of research has to resign herself to the fact that it will date quicker than will most theoretical work. It can have a more immediate payoff, however, if experimental systems subsequently go into production, and it can yield insights that can immediately be used in the next project. The study discussed in this chapter ultimately provided some early proof-of-concept in support of automating or partially automating functions such as appraisal, accessioning or even potentially description and retrieval. To be taken further, however, it would have needed to have been a part of a series of ongoing, cumulative studies. Experimental prototypes such as the one described here that was developed using a test collection might then be implemented and tested with a real-life system, and subsequently, perhaps with multiple real-life systems with similar or different characteristics in order to isolate what might be effective and what not. Such a trajectory of studies was beyond the scope of work described here and really has yet to substantively occur in archival systems design and development.

Since conducting this study, the fields of knowledge extraction and data mining have advanced considerably although expert systems are still widely used in business, medicine and certain other sectors. The research community has grown impatient with the slow progress in the field of artificial intelligence research out of which this method emerged and now favours other forms of community knowledge extraction such as ontological modelling.[59] Information communications and recordkeeping technologies

have also evolved quite dramatically, with administrative and personal activities being conducted in a highly networked and mobile fashion in a constant accumulation of meanings and metadata (see, again, Acker's chapter in this volume). As a result, there have been serious challenges made to the continuing necessity and utility of appraisal, a solid part of the archival field for over a century. The notion of the self-describing record and intensified use of information retrieval and data mining techniques, as with the Twitter Archive, similarly challenge the traditional parameters envisaged around description.

At the same time, however, as the University of Oregon case illustrates, there is an increasingly demonstrable need for more automated assistance with evaluating, processing and retrieving exceedingly high volumes of documentation generated by computer-mediated communications in both recordkeeping and other documentary contexts. The archival field is in almost the direct inverse of the situation in which information retrieval or indeed the library management software finds themselves: there has been a lot of theorizing, a small amount of modelling and rules production, and an even smaller amount of actual systems development.[60] It remains unknown which aspects of IR might transfer into archival science (even bibliometric norms do not appear to fit, and precision and recall have been demonstrated to be not necessarily useful measures), and IR and systems design, which are cumulative by nature, have no track record to build upon. And yet systems development research is going to be essential if archives are to work with CMC and other forms of digital materials, and if archivists choose not to engage with it, contemporary examples would suggest that there are plenty of other communities who are desirous to do so, and who may not have the same sensibilities as archivists. A robust trajectory of archival systems development needs not only researchers with sound technological and IR expertise, but also models, hypotheses, unambiguous definitions and schematics, experimental and live testbeds and test collections, and a culture of evaluation and refinement.

References

Abraham, Ajith, Aboul-Ella Hassanien and Václav Suásel, eds. *Computational Social Network Analysis: Trends, Tools and Research Advances*. London: Springer-Verlag, 2011.

Anderson, Kimberly. *Appraisal Learning Networks: How University Archivists Learn to Appraise Through Social Interaction*, doctoral dissertation. Los Angeles: University of California Los Angeles, 2011.

Archivists' Toolkit, http://archiviststoolkit.org/.

Arnold, Timothy Jason. "Digital Curation in the Age of Twitter: Curating a Collection of Tweets from the Egyptian Revolution," paper presented at the Council of State Archives and Society of American Archivists Joint Annual Meeting, New Orleans, August 2013.

Bailey, Catherine. "Archival Theory and Electronic Records," *Archivaria* 29 (Winter 1989-90): 180-196.

Bearman, David A. ed. *Archival Management of Electronic Records*. Pittsburgh, PA: Archives and Museum Informatics Technical Report no. 13, 1991.

Bearman, David A. “Archives in the Post-Custodial Age,” *Archival Management of Electronic Records, Archives and Museum Informatics Technical Report* No. 13. Pittsburgh, PA: Archives and Museum Informatics, 1991.

Bearman, David A. “Diplomatics, Weberian Bureaucracy, and the Management of Electronic Records in Europe and America,” *The American Archivist* 55 (1992): 168–181.

Bearman, David A. *Electronic Evidence: Strategies for Managing Records in Contemporary Organizations*. Pittsburgh, PA: Archives and Museum Informatics, 1994.

Bearman, David A. “Electronic Mail,” *Electronic Evidence: Strategies for Managing Records in Contemporary Organizations*. Pittsburgh, PA: Archives and Museum Informatics, 1994.

Bearman, David A. “Electronic Records Guidelines,” in *Electronic Evidence: Strategies for Managing Records in Contemporary Organizations*. Pittsburgh, PA: Archives and Museum Informatics, 1994.

Bearman, David A. *Electronic Records Guidelines: A Manual for Policy Development and Implementation*. Pittsburgh, PA: Archives and Museum Informatics, 1990.

Bearman, David A. “The Implications of Armstrong v. Executive Office of the President for the Archival Management of Electronic Records,” *The American Archivist* 56 (1993): 674–689.

Bearman, David A. “An Indefensible Bastion: Archives as a Repository in the Electronic Age,” in David Bearman, ed. *Archival Management of Electronic Records*. Pittsburgh: Archives & Museum Informatics, 1991, pp. 14-24.

Bearman, David A. “Record-keeping Systems,” *Archivaria* 36 (1993): 16-37.

Bearman, David A. "Towards a Reference Model for Business Acceptable Communications,” unpublished paper, December 6, 1994.

Bikson, Tora. “Research on Electronic Information Environments: Prospects and Problems,” paper presented 24 January 1991 at the Working Meeting on Research Issues in Electronic Records. Washington, D.C.: National Historical Publications and Records Commission, 1991.

BitCurator, http://www.bitcurator.net/.

Blair, David C. *Languages and Representation in Information Retrieval (*New York: Elsevier Science, 1990).

Boles, Frank. "SAA Should Say Something!" Society of American Archivists press release (February 2015), http://www2.archivists.org/news/2015/saa-should-say-something.

Borgman, Christine L. ed. “Bibliometrics and Scholarly Communications,” special issue of *Communication Research* (October 1989).

Burckel. Nicholas C. “The Expanding Role of A College or University Archives,” *Midwestern Archivist* 1 (1976): 16-32.

Consultative Committee for Space Data Systems/International Organization for Standardization. *Space Data and Information Transfer System: Open Archival Information System: Reference Model*. Geneva, Switzerland: International Organization for Standardization, 1999.

Cook, Terry. *The Archival Appraisal of Records Containing Personal Information: A RAMP Study with Guidelines*. Paris: UNESCO, April 1991.

Cook, Terry. “Easier to Byte, Harder to Chew: The Second Generation of Electronic Records Archives,” *Archivaria* 33 (1992): 202–216.
Cook, Terry. “Electronic Records, Paper Minds: The Revolution in Information Management and Archives in the Post-custodial and Post-modern Era,” *Archives and Manuscripts* 22 (1994): 300–329.
Cook, Terry. “It’s 10 O’Clock—Do You Know Where Your Data Are?” *Technology Review* 98 (1995): 48–53.
Cook, Terry and Eldon Frost. “The Electronic Records Archival Programme at the National Archives of Canada: Evolution and Critical Factors of Success,” in Margaret Hedstrom, ed, *Electronic Records Management Program Strategies: Archives and Museum Informatics Technical Report No. 18*. Pittsburgh, PA: Archives and Museum Informatics, 1993, pp. 38–47.
Cox, Richard J., “American Archival Literature: Expanding Horizons and Continuing Needs, 1901-1987,” *The American Archivist* 50 (1987): 306-323.
Cox, Richard J. “Re-discovering the Archival Mission: The Recordkeeping Functional Requirements Project at the University of Pittsburgh, A Progress Report. *Archives and Museum Informatics* 8 (1994): 279–300.
Cox, Richard J. “Searching for Authority: Archivists and Electronic Records in the New World at the *fin-de-siécle*,” *First Monday* 5 nos. 1-3 (January 2000).
Cox, Richard J., and Wendy Duff. “Warrant and the Definitions of Electronic Records: Questions Arising from the Pittsburgh Project. *Archives and Museum Informatics* 11 (1997): 223–231.
De Bellis, Nicola. *Bibliometrics and Citation Analysis: From the Science Citation Index to Cybermetrics*. Chicago: Scarecrow Press, 2009.
Dietz, Diane. "Former UO Archivist Describes ‘Humiliating’ Dismissal: He Says University Leaders Saw Him as a Scapegoat after a Controversial Records Release," *The Register-Guard* (Sunday, April 19th, 2015), http://registerguard.com/rg/news/local/32990068-75/former-uo-archivist-james-fox-tells-his-side-of-his-dismissal.html.csp**.**
Duranti, Luciana, Terry Eastwood, and Heather MacNeil. *Preservation of the Integrity of Electronic Records*. Dordrecht, The Netherlands: Kluwer Academic, 2002.
Duranti, Luciana and Heather MacNeil. “The Protection of the Integrity of Electronic Records: An Overview of the UBC-MAS Research Project,” *Archivaria 42* (1996): 46–67.
Duranti, Luciana, and Heather MacNeil. “Protecting Electronic Evidence: A Third Progress Report on a Research Study and its Methodology,” *Archivi & Computer* 6 no.5 (1996): 343–404.
Gilliland, Anne J. “Archival Appraisal: Practising on Shifting Sands,” in *Archives and Recordkeeping: Theory Into Practice*, Patricia Whatley and Caroline Brown, eds. London: Facet Press, 2014.
Gilliland, Anne. “Reflections on the Value of Metadata Archaeology for Recordkeeping in a Global, Digital World,” *Journal of the Society of Archivists* 32 no.1 (April 2011): 97-112.
Gilliland, Anne J. and Sue McKemmish. "Archival and Recordkeeping Research: Past, Present and Future," in *Research Methods: Information Management, Systems,*

and Contexts Kirsty Williamson, ed. Prahran: Tilde University Press, 2012, pp.80-112.

Gilliland-Swetland, Anne J. "Archivy and the Computer: A Citation Analysis of North American Periodical Articles," *Archival Issues* 17 no. 2 (1992): 95-112.

Gilliland-Swetland, Anne J. *Policy and Politics: Electronic Communications and Electronic Culture at The University of Michigan, A Case Study in the Management of Electronic Mail*, in Society of American Archivists case study series *Archival Administration of Electronic Records and the Use of New Technologies in Archives*. Chicago, IL: Society of American Archivists, 1996.

Gilliland-Swetland. Anne J. and Carol Hughes. "Enhancing Archival Description for Public Computer Conferences of Historical Value: An Exploratory Study," *The American Archivist* 55 no.2 (Spring 1992): 316-330.

Gilliland-Swetland, Anne J. and Gregory T. Kinney. "Uses of Electronic Communications to Document an Academic Community: A Research Report," *Archivaria* 38 (Fall 1994): 79-96.

Gleick, James. "Roving Librarians of the Twitterverse," *New York Review of Books Blog*, January 16, 2013, www.nybooks_com_blogs_nyrblog_2013_jan_16_librarians_twi.pdf

Hedstrom, Margaret. "Understanding Electronic Incunabula: A Framework for Research on Electronic Records," paper presented 24 January 1991 at the Working Meeting on Research Issues in Electronic Records. Washington, D.C.: National Historical Publications and Records Commission, 1991.

Kalinichenko, Leonid, Michele Missikoff, Federica Schiappelli, Nikolay Skvortsov. "Ontological Modeling," *Proceedings of the 5th Russian Conference on Digital Libraries RCDL2003, St.-Petersburg, Russia, 2003*, citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.101.5714.pdf

Library of Congress. "Update on the Twitter Archive at the Library of Congress," January 2013.

McKemmish, Sue and Frank Upward. "Somewhere Beyond Custody," *Archives and Manuscripts* 22 no.1 (1994): 138-149.

McQuail, Denis. *McQuail's Mass Communication Theory*. 5th ed. (London: Sage Publications, 2005).

Morozov, Evgeny. "Facebook and Twitter are Just Places Revolutionaries Go," *The Guardian* (7 March 2011).

Nunamaker, Jay F. Jr. and Minder Chen. "Systems Development in Information Systems Research," *Proceedings of the Twenty-Third Annual Hawaii International Conference on Systems Sciences, IEEE,* 1990, pp. 631-640

O'Shea, Greg and David Roberts. "Living in a Digital World: Recognizing the Electronic and Post-custodial Realities, *Archives and Manuscripts* 24 no.2 (1996): 286-311.

Pao, Miranda Lee. *Concepts of Information Retrieval*. Englewood: Libraries Unlimited, 1989.

Pinheiro, Carlos A.R. *Social Network Analysis in Telecommunications*. New York: John Wiley and Sons, 2011.

Pinheiro, Carlos A.R. and Marcus Helfert, eds. *Exploratory Analysis in Dynamic Social Networks*. Hong Kong: iConcept Press, 2012.

Rao, Aluri and Donald E. Riggs. “Applications of Expert Systems in Libraries,” *Advances in Library Automation and Networking* 2 (1988): 1-43.
Salton, Gerard and Chris Buckley. "Global Text Matching for Information Retrieval," *Science* 253 (1991): 1012-1015.
Samuels, Helen W. "Who Controls the Past?" *The American Archivist* 49 no.2 (Spring 1986): 109–124.
Samuels, Helen W. *Varsity Letters: Documenting Modern Colleges and Universities*. Chicago: Society of American Archivists and Scarecrow Press, 1992.
Thurlow, C., L. Lengel, and A. Tomic. *Computer Mediated Communication: Social Interaction and the Internet*. London: Sage, 2004.
Upward, Frank, Sue McKemmish and Barbara Reed. “Archivists and Changing Social and Information Spaces: A Continuum Approach to Recordkeeping and Archiving in Online Cultures,” *Archivaria* 72 (Fall 2011): 197-238.
Vickery, Alina and Helen Brooks. *Expert System for Referral*. Library and Information Science Research Report No. 66, British Library, 1988.

[1] The author would like to thank Amelia Acker for her close reading and insightful comments on drafts of this paper.

[2] Computer-mediated Communications (CMC) refers to any communicative transaction, such as email, text messaging, and instant messaging, that employs two or more networked computers. In recent years, the term has increasingly been used in the context of research into the social effects and implications of such technologically-mediated communication. See Denis McQuail, *McQuail's Mass Communication Theory*. 5th ed. London: Sage Publications, 2005 and C. Thurlow, L. Lengel, and A. Tomic, *Computer Mediated Communication: Social Interaction and the Internet*. London: Sage, 2004.

[3] Diane Dietz, "Former UO Archivist Describes 'Humiliating' Dismissal: He Says University Leaders Saw Him as a Scapegoat after a Controversial Records Release," *The Register-Guard* (Sunday, April 19th, 2015), http://registerguard.com/rg/news/local/32990068-75/former-uo-archivist-james-fox-tells-his-side-of-his-dismissal.html.csp.

[4] Nunamaker and Chen described systems development as both a research methodology and a research domain, especially as used in the field of information systems. As a methodology, it falls under applied science and "belongs to the engineering, developmental and formulative type of research" (p.632). They argued that the systems building process, especially one that includes software development, moves through the following phases: construction of a conceptual framework; development of systems architecture; analysis and design of the system; building the system; and observing and evaluating the system. Jay F. Nunamaker, Jr. and Minder Chen, "Systems Development in Information Systems Research," *Proceedings of the Twenty-Third Annual Hawaii International Conference on Systems Sciences, IEEE*, 1990, pp. 631-640. More recent developments in systems development, such as those described in by Joanne Evans in her chapter in this volume, integrate more reflexive and even qualitative approaches for a less technologically deterministic methodological approach.

[5] See Tora Bikson, "Research on Electronic Information Environments: Prospects and Problems" and Margaret Hedstrom, "Understanding Electronic Incunabula: A Framework for Research on Electronic Records," papers presented 24 January 1991 at the Working Meeting on Research Issues in Electronic Records. Washington, D.C.: National Historical Publications and Records Commission, 1991.

[6] *Armstrong v. Executive Office of the President*. See David A. Bearman. "The Implications of Armstrong v. Executive Office of the President for the Archival Management of Electronic Records," *The American Archivist* 56 (1993): 674–689.

[7] Both individuals emphasised how electronic records could serve as digital evidence of decision-making and transactions occurring in the course of organisational activity, and looked for ways to identify how and ensure that recordkeeping systems created such evidence. See Terry Cook, "Easier to Byte, Harder to Chew: The Second Generation of Electronic Records Archives," *Archivaria* 33 (1992): 202–216; "Electronic Records, Paper Minds: The Revolution in Information Management and Archives in the Post-custodial and Post-modern Era," *Archives and Manuscripts* 22 (1994): 300–329; "It's 10 O'Clock—Do You Know Where Your Data Are?" *Technology Review* 98 (1995): 48–53; Terry Cook and Eldon Frost, "The Electronic Records Archival Programme at the National Archives of Canada: Evolution and Critical Factors of Success," in Margaret Hedstrom, ed, *Electronic Records Management Program Strategies: Archives and Museum Informatics Technical Report No. 18*. Pittsburgh, PA: Archives and Museum Informatics, 1993, pp. 38–47; David A. Bearman, *Electronic Records Guidelines: A Manual for Policy Development and Implementation*. Pittsburgh, PA: Archives and Museum Informatics, 1990; (ed.) *Archival Management of Electronic Records*. Pittsburgh, PA: Archives and Museum Informatics Technical Report no. 13, 1991; "Archives in the Post-Custodial Age," ibid.; "An Indefensible Bastion: Archives as a Repository in the Electronic Age," in *Archival Management*," pp.14-24; "Diplomatics, Weberian Bureaucracy, and the Management of Electronic Records in Europe and America," *The American Archivist* 55 (1992): 168–181; "Record-keeping Systems," *Archivaria* 36 (1993): 16-37; *Electronic Evidence: Strategies for Managing Records in Contemporary Organizations*. Pittsburgh, PA: Archives and Museum Informatics, 1994; Sue McKemmish and Frank Upward, "Somewhere Beyond Custody," *Archives and Manuscripts* 22 no.1 (1994): 138-149; Greg O'Shea and David Roberts, "Living in a Digital World: Recognizing the Electronic and Post-custodial Realities, *Archives and Manuscripts* 24 no.2 (1996): 286-311.

[8] Anne J. Gilliland and Sue McKemmish, "Archival and Recordkeeping Research: Past, Present and Future," in *Research Methods: Information Management, Systems, and Contexts* Kirsty Williamson, ed. Prahran: Tilde University Press, 2012, pp.80-112.

[9] In some ways, it could be argued that the introduction of first the MARC Archival and Manuscripts Control Format and then Encoded Archival Description in the U.S. did indeed have such an effect on the descriptive practices of C&U archives.

[10] Anne J. Gilliland, "Reflections on the Value of Metadata Archaeology for Recordkeeping in a Global, Digital World,"

Journal of the Society of Archivists 32 no.1 (April 2011): 97-112.

[11] Anne J. Gilliland-Swetland and Carol Hughes, “Enhancing Archival Description for Public Computer Conferences of Historical Value: An Exploratory Study,” *The American Archivist* 55 no.2 (Spring 1992): 316-330; Gilliland-Swetland, Anne J. and Gregory T. Kinney, “Uses of Electronic Communications to Document an Academic Community: A Research Report,” *Archivaria* 38 (Fall 1994): 79-96.

[12] Helen W. Samuels, *Varsity Letters: Documenting Modern Colleges and Universities*. Chicago: Society of American Archivists and Scarecrow Press, 1992.

[13] Richard J. Cox, “Re-discovering the Archival Mission: The Recordkeeping Functional Requirements Project at the University of Pittsburgh, A Progress Report. *Archives and Museum Informatics* 8 (1994): 279–300; Richard J. Cox and Wendy Duff. “Warrant and the Definitions of Electronic Records: Questions Arising from the Pittsburgh Project," *Archives and Museum Informatics* 11 (1997): 223–231.

[14] Luciana Duranti, Terry Eastwood, and Heather MacNeil, *Preservation of the Integrity of Electronic Records*. Dordrecht, The Netherlands: Kluwer Academic, 2002; Luciana Duranti and Heather MacNeil, “The Protection of the Integrity of Electronic Records: An Overview of the UBC-MAS Research Project,” *Archivaria 42* (1996): 46–67, and “Protecting Electronic Evidence: A Third Progress Report on a Research Study and its Methodology,” *Archivi & Computer* 6 no.5 (1996): 343–404.

[15] Examples would include the National Archives and Records Administration and the State Archives and Records Administration of New York.

[16] For further discussion of these and other management issues at the time, see Anne J. Gilliland-Swetland, *Policy and Politics: Electronic Communications and Electronic Culture at The University of Michigan, A Case Study in the Management of Electronic Mail*, in Society of American Archivists case study series *Archival Administration of Electronic Records and the Use of New Technologies in Archives*. Chicago: Society of American Archivists, 1996.

[17] Communications researchers, especially those who were sociometricians and bibliometricians were particularly interested in this area.

[18] A documentation strategy is an appraisal methodology that, “consists of four activities: 1. choosing and defining the topic to be documented, 2. selecting the advisors and establishing the site for the strategy, 3. structuring the inquiry and examining the form and substance of the available documentation, and 4. selecting and placing the documentation.” Helen W. Samuels, "Who Controls the Past?" *The American Archivist* 49 no.2 (Spring 1986): 116.

[19] Terry Cook, *The Archival Appraisal of Records Containing Personal Information: A RAMP Study with Guidelines*. Paris: UNESCO, April 1991: 4.

[20] “A modern, complex, information-rich society requires that archivists reexamine their role as selectors. The changing structure of modern institutions and the use of sophisticated technologies have altered the nature of records, and only a small portion of the vast documentation can be kept. Archivists are challenged to select a lasting record, but they lack techniques to support this decision-making. Documentation strategies are proposed to respond to these problems.” Samuels, "Who Controls the Past?", 109. In this, Samuels was echoing another university archivist, Nicholas Burckel, who wrote that: “Another collecting focus [of] college and university archives should be the intellectual and cultural atmosphere which the university engenders. This can hardly be determined from a look at transcripts of college catalogs.” See Nicholas C. Burckel, “The Expanding Role of A College or University Archives,” *Midwestern Archivist* 1 (1976): 5.

[21] Samuels, *Varsity Letters*.

[22] Gilliland and Kinney, “Electronic Communications.”

[23] See, in particular, the collection of essays by authors such as Christine Borgman, Leah Lievrouw, Belver Griffith, and William Paisley contained in “Bibliometrics and Scholarly Communications,” a special issue of *Communication Research* (October 1989) edited by Christine L. Borgman.

[24] In this context, Bearman defined records as:

> at one and the same time the carriers, products, and documentation, of transactions. Transactions are, by definition, communicated from one person to another, from a person to a store of information (filing cabinet, computer database) available to another person at a later time, or from a store of information to a person or

> another computer. As such, transactions must leave the mind or computer memory in which they are created, or must be used by a person with access to the same computer memory. The transaction record must be conveyed across a software layer, and typically across a number of hardware switches.

David A. Bearman, "Towards a Reference Model for Business Acceptable Communications," unpublished paper, December 6, 1994.

[25] David A. Bearman, "Electronic Records Guidelines," in *Electronic Evidence: Strategies for Managing Records in Contemporary Organizations*. Pittsburgh, PA: Archives and Museum Informatics, 1994, p. 100.

[26] These user concerns had been previously identified by the Bentley Conferencing Appraisal Project.

[27] An X.400 system.

[28] This was an approach alluded to by Terry Cook, when he stated that, "Records centre storage can be used to gain distance and perspective in making appraisal decisions, especially for those series of records involving public issues or government functions which are controversial, hotly debated in public forums, and emotion laden for many citizens (including archivists) at the time of their occurrence." Cook, *Archival Appraisal*, 81.

[29] Anne J. Gilliland-Swetland, "Archivy and the Computer: A Citation Analysis of North American Periodical Articles," *Archival Issues* 17 no.2 (1992): 95-112.

[30] See Nicola De Bellis. *Bibliometrics and Citation Analysis: From the Science Citation Index to Cybermetrics*. Chicago: Scarecrow Press, 2009.

[31] This facet of bibliometrics is evidenced in the context of this study in that it built on prior applications of bibliometrics, and Kimberly Anderson's study (discussed in her chapter in this volume), in turn, applies some of the rules used in the study discussed in this chapter.

[32] The periodicals examined were the *American Archivist, Prologue, Archival Issues* (formerly the *Midwestern Archivist*), *Provenance* (formerly *Georgia Archive*), *Archivaria*, and the *Public Historian*, which at the time, according to Richard Cox: " . . . should be, essential reading for any archivist." These journals, with the exception of *Prologue*, which was an official publication of the National Archives and Records Administration, were published by major archival associations in the U.S. As such, most of the journals were distributed free to members of those organisations and also were sent by subscription to many archival and library repositories. See Richard J. Cox, "American Archival Literature: Expanding Horizons and Continuing Needs, 1901-1987," *The American Archivist* 50 (1987): 317.

[33] In fact, subsequent bibliometric research conducted by Richard Cox and by Kimberly Anderson both drew upon the procedures laid out in this study. See Richard J. Cox, "Searching for Authority: Archivists and Electronic Records in the New World at the *fin-de-siécle*," *First Monday* 5 nos. 1-3 (January 2000). Available: http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/721/630; and Kimberly Anderson. *Appraisal Learning Networks: How University Archivists Learn to Appraise Through Social Interaction*, doctoral dissertation. Los Angeles: University of California Los Angeles, 2011.

[34] A simple approach to identifying which articles and other texts to analyse would have been to select a well known library and information science or humanities indexing, abstracting or citation resource and then to cull all those items covered in the last 21 years that were indexed under the professional terminology "archival appraisal." This was not feasible, however, since archival science to a certain extent was located on the periphery of each of the coverage areas of such resources and in no case were all of the journals to be used in the citation analysis indexed and/or abstracted by any one resource. Moreover, in the instances where one of the journals was indexed or abstracted, the subject headings used were not sufficiently accurate regarding the archival definition of appraisal to be useful for this citation analysis. Using titles cataloged in OCLC, monographs were selected for analysis, therefore, where relevant word stems (e.g., "apprais*", "sampl*" or "select*"), terms (e.g., "valuation" or "assessment") or phrases (e.g., "documentation strategy" or "documentation plan*") appeared in either the title statement, the index terms assigned by Library of Congress Subject Headings, or both. In the case of periodicals, each issue was physically examined and for an article to be selected for analysis, the word stems "apprais*" and/or "select*;" and/or either of the phrases "documentation strategy" or "documentation plan*" had to appear three or more times in any or all of the title, abstract, or first or last paragraphs. These words stems and the frequency of their occurrence were chosen to try to avoid false drops in the form of articles or texts relating to the appraisal of manuscript materials for monetary value, an entirely different area of study. These rules were based on those used by the author for a previous citation analysis of archival literature. See Gilliland-Swetland, "Archivy and the Computer."

[35] The analysis of citations and references contained in the literature applied the following rules:

a) References contained in footnotes, endnotes, and bibliographies were all analysed for both monographs, and monographic and journal articles. Since the journals varied widely in the type of work that they presented as an article (e.g., review articles, commentaries, case studies, author responses, proceedings of meetings, and news items) a working definition of what constituted an article within the scope of this study excluded news notes, letters to the editor, and book reviews but included all the other types listed above;
b) The number of citations made to each individual author was counted;
c) Citations to multiple works by the same author were counted;
d) Multiple citations within an analysed publication to the same work were counted only once;
e) In cases where citations are to works with multiple authors, each author was counted separately;
f) Citations to non-North American authors were included under the assumption that they were exerting an influence on American thinking about appraisal.
This work was all done by hand, but today more of this can be automated because so much is now in digital form. However, not all the publications are indexed by the leading indexing or citation services and indeed some that used to be have been dropped in recent years. This constitutes a major logistical limitation to using bibliometrics in archival research.

[36] In the initial proposal for this research and consistent with bibliographic citation norms for scientific literature, I had expected that living authors who were included in the top 50% of citations would be those selected as experts. However, the resulting bibliometric data did not match the expected norms, perhaps because archivists during the time period covered, coming from a strong historical disciplinary approach, might not have used footnotes and references in the same way as did scientific authors; or simply because they tended to publish less and the literature was consequently more scattered.

[37] An expert "shell" is a software package that facilitates the development of an expert system. Shells provide a minimum of an inference engine, interface design capabilities, and a mechanism for developing and integrating rules. The actual content of the system is added by the systems developer or knowledge engineer.

[38] In terms of expert system capabilities, I had sought a minimum of the following: ability to work with frames (a structure for representing facts or data, procedures and default values that was originally derived from semantic networks and that can be linked together like building blocks) and inheritance; ability to handle a mid-sized rule-base; and ability to function as a front-end to an electronic communications system or file of communications. Whether the system would need to be able to work with semantic networks would depend upon whether rules were developed that might require such a level of natural language processing.

[39] See Aluri Rao and Donald E. Riggs, "Applications of Expert Systems in Libraries," *Advances in Library Automation and Networking* 2 (1988): 1-43.

[40] The most detailed discussion of evaluation criteria found was contained in Vickery and addressed to the library reference environment:

- System power - does the system exhibit intelligent behaviour in performing complex tasks?
- How robust was the system with imprecise or incomplete data?
- How flexible was the system?
- System response time.
- Transparency of the system, and effectiveness of the user interface and explanation facility.
- Does the system perform equal to a human?
- Hardware and software performance.
- Ease of use.
- Completeness and consistency of the knowledge base.
- Comparison with the same library service when performed manually.

Unfortunately, many of these evaluation criteria proved either to be impossible to implement because of system and data limitations, or would have yielded results that would have been hard to interpret because of confounding variables. Instead I used the evaluation mechanisms outlined below, and while they drew upon some of Vickery's criteria, they were tailored more closely to the archival environment and the task to be performed in this instance. See Alina Vickery and Helen Brooks, *Expert System for Referral*. Library and Information Science Research Report No. 66, British Library, 1988.

[41] Researchers at the Massachusetts Institute of Technology (MIT) have been working in this area for many years, and as already noted, some of that research influenced the study discussed here. Most recently, in July 2013, MIT researchers announced the development of Immersion, an application that offers many qualities that archivists might consider when developing digital tools to support how their own users might approach "reading" and analysing archived CMC: "Immersion is an invitation to dive into the history of your email life in a platform that offers you the

safety of knowing that you can always delete your data. Just like a cubist painting, Immersion presents users with a number of different perspectives of their email data. It provides a tool for self-reflection at a time where the zeitgeist is one of self-promotion. It provides an artistic representation that exists only in the presence of the visitor. It helps explore privacy by showing users data that they have already shared with others. Finally, it presents users wanting to be more strategic with their professional interactions, with a map to plan more effectively who they connect with." See https://immersion.media.mit.edu/

[42] It is worth noting, however, that despite what my own research indicates about the evidentiary importance of CMC metadata, and the deep concerns expressed by many archivists, the U.S. National Archives has asserted that "the data being collected by NSA is classified as "raw signal intelligence." Both congressional legislation and administrative order define "raw signal intelligence" as a nonpermanent, federal record. It will be destroyed." Frank Boles, "SAA Should Say Something!" Society of American Archivists press release (February 2015), http://www2.archivists.org/news/2015/saa-should-say-something.

[43] Catherine Bailey, "Archival Theory and Electronic Records," *Archivaria* 29 (Winter 1989-90): 73.

[44] Library of Congress, "Update," p.1.

[45] Library of Congress. "Update on the Twitter Archive at the Library of Congress," January 2013, p.5.

[46] See, for example, James Gleick, "Roving Librarians of the Twitterverse," *New York Review of Books Blog*, January 16, 2013, www.nybooks_com_blogs_nyrblog_2013_jan_16_librarians_twi.pdf; Evgeny Morozov, "Facebook and Twitter are Just Places Revolutionaries Go," *The Guardian*, 7 March 2011.

[47] The process of determining the value and thus the disposition of records based upon their current administrative, legal, and fiscal use; their evidential and informational value; their arrangement and condition, their intrinsic value; and their relationship to other records. A broader definition offered by documentation strategists is that of any selection activity that enables archivists to identify recorded information that has enduring value.

[48] The term "ingestion," which is increasingly used instead of the archival term "accession" in relation to the acquisition and initial processing of digital materials, reflects the influence of the language used in the Open Archival Information System (OAIS) Reference Model, first developed in 1997 by the space science community as a high level model for the "archiving" of the massive quantities of digital data being generated in the space science community through such activities as remote sensing via satellite. See the Consultative Committee for Space Data Systems/International Organization for Standardization. *Space Data and Information Transfer System: Open Archival Information System: Reference Model*. Geneva, Switzerland: International Organization for Standardization, 1999.

[49] Library of Congress. "Update," p.1.

[50] Library of Congress, "Update," p.1.

[51] For example, Gnip is required to protect deleted tweets from public disclosure. In a paper world, materials that were intended to be deleted would likely not even reach the archives, but if they did, archivists would make a decision about their disposition as part of physically processing the entire accumulation or collection. In a digital world, however, the traces of deleted material, if not the actual deleted material (as in the case of Twitter) often persist, and the approach taken by the Twitter Archive is not to attempt to eliminate them before or after acquiring them, but instead to render them publicly irretrievable through the retrieval mechanisms being developed for use with the Archive.

[52] Archivists, in conducting appraisal and description, which might be viewed as the parallel archival activities to library selection and cataloging, are concerned about collectively assessing and elucidating the value of the materials in hand as evidence of bureaucratic, community and personal activity as well as for their informational value.

[53] David A. Bearman, "Archives in the Post-Custodial Age," *Archival Management of Electronic Records, Archives and Museum Informatics Technical Report* No. 13. Pittsburgh, PA: Archives and Museum Informatics, 1991. Gerard Salton and Chris Buckley developed a system around the same time that used flexible text matching procedures to retrieve documents from large text collections with unrestricted subject matter. Salton and Buckley's method broke texts into semantically homogeneous excerpts such as paragraphs or sentences; uses a standard automatic indexing process to assign natural language weighted content identifiers; and then detects similarities between particular text items, or between text items and search requests, by comparing term vectors at various levels of detail. The authors gave several examples of how the system works, intellectually and mathematically, one of which was testing recall and precision in retrieval from a non-bibliographic collection of electronic mail. See Gerard Salton and Chris Buckley, "Global Text Matching for Information Retrieval," *Science* 253 (1991): 1012-1015.

[54] Frank Upward, Sue McKemmish and Barbara Reed, “Archivists and Changing Social and Information Spaces: A Continuum Approach to Recordkeeping and Archiving in Online Cultures,” *Archivaria* 72 (Fall 2011): 197-238.

[55] Social network analysis is primarily a quantitative method that is concerned with social and cultural aspects and effects of virtual social structures. It typically employs two categories of methods to look at relationships between network members such as their positions, strength and clustering and diversity of their connections: network data collection (either socio-centric or ego-centric) that looks at the strength, confirmation and multiplicity of networked relationships of groups or individuals; and network data visualisation, that uses various visualisation methods alone or in combination such as maps and matrices. See Ajith Abraham, Aboul-Ella Hassanien and Václav Suásel, eds., *Computational Social Network Analysis: Trends, Tools and Research Advances*. London: Springer-Verlag, 2011; Carlos A.R. Pinheiro, *Social Network Analysis in Telecommunications*. John Wiley and Sons, 2011; and Carlos A.R. Pinheiro and Marcus Helfert, eds., *Exploratory Analysis in Dynamic Social Networks*. iConcept Press, 2012.

[56] In his research relating to the use of Twitter in Egypt, Arnold discusses the development of an API that allows the curation of information on individual users who an archivist might be interested in following (such as “elite” or professional tweeters), using the user’s profile page, and allowing the researcher to track such aspects as the user’s bio, who s/he is following, and who is following him/her. For search terms, it is possible to can use common and niche hashtags, threads. Arnold says that methods can be qualitative, e.g., a researcher setting up his/her own account and embedding him/herself in conversations, or qualitative, e.g., counting tweets, retweets and mentions, and these can be analysed respectively using content or statistical analysis software. Scraping software is also available. Arnold also notes that a research might examine language issues, e.g., colloquial or formal, vernacular or standardised. See Timothy Jason Arnold, “Digital Curation in the Age of Twitter: Curating a Collection of Tweets from the Egyptian Revolution,” paper presented at the Council of State Archives and Society of American Archivists Joint Annual Meeting, New Orleans, August 2013.
Another approach developed by George Washington University Libraries that can be used for archival acquisition and analysis purposes is Social Feed Manager, which manages rules and streams from social data sources. See https://github.com/gwu-libraries/social-feed-manager.

[57] Jerome McDonough, comments made at the Digital Preservation Symposium, Ann Arbor, June 25-27, 2011.

[58] Reflecting critically upon the interpretative frameworks that are brought to bear in systems development and how they have been constructed, as well as upon the systems developers’ own actions.

[59] “Generally an ontology can be defined as a linguistic artifact that defines a shared vocabulary of basic concepts for discourse about a piece of reality (subject domain) and specifies what precisely those concepts mean. As such, ontologies provide the basis for semantic modeling of subject domain, information integration, and communication in the domain.” See Leonid Kalinichenko, Michele Missikoff, Federica Schiappelli, Nikolay Skvortsov, “Ontological Modeling,” *Proceedings of the 5th Russian Conference on Digital Libraries RCDL2003, St.-Petersburg, Russia, 2003*, citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.101.5714.pdf

[60] Currently the two most prominent tools in the United States are the Archivists' Toolkit, an archival processing tool which "supports accessioning and describing archival materials; establishing names and subjects associated with archival materials, including the names of donors; managing locations for the materials; and exporting EAD finding aids, MARCXML records, and METS, MODS and Dublin Core records. Future functionality will be built to support repository user/resource use information, appraisal for archival materials, expressing and managing rights information, and interoperability with user authentication systems," http://archiviststoolkit.org/node/96; and BitCurator, a digital forensics tool designed for archives, libraries and museums which can "capture bit-for-bit copies of data contained on digital storage devices, scan digital holdings for sensitive information, generate technical metadata reports detailing the content of digital media, and more", http://mith.umd.edu/research/project/bitcurator/. Both developed out of federally funded research and development projects and while they are certainly important steps in the right direction neither can yet adequately address the kinds of concerns and scope raised by the University of Oregon case. The ongoing BitCurator Access project directed by Christopher A. Lee at the University of North Carolina, Chapel Hill promises to address more of the issues associated with sensitive content and metadata. It aims to develop "open-source software that supports the provision of access to disk images through three exploratory approaches: (1) building tools to support web-based services, (2) enabling the export of file systems and associated metadata, (3) and the use of emulation environments. Also closely associated with these access goals is redaction. BitCurator Access will develop tools to redact files, file system metadata, and targeted bitstreams within disks or directories," http://www.bitcurator.net/bitcurator-access-people/.