# Computer-Assisted Appraisal and Selection of Archival Materials

Christopher A. Lee
University of North Carolina
Chapel Hill, USA
callee [@ils.unc.edu]

***Abstract*— Despite progress on various technologies to support both digital preservation and description of archival materials, we have still seen relatively little progress on software support for the core activities of selection and appraisal. There are two considerations that make selection and appraisal of digital materials substantially different from selection and appraisal of analog materials: that digital materials exist at multiple levels of representation and that they are directly machine readable. There are great opportunities to better assist selection and appraisal of digital materials, including use of digital forensics tools, natural language processing, and machine learning.**

***Keywords—appraisal, selection, digital curation, archival science, natural language processing, digital forensics***

## I. INTRODUCTION AND RATIONALE

Despite a couple decades of progress on various technologies to support both digital preservation and description of archival materials, we have still seen relatively little progress on software support for the core activities of selection and appraisal. There are two considerations that make selection and appraisal of digital materials substantially different from selection and appraisal of analog materials.

The first consideration is that digital materials exist at multiple levels of representation. If records are "persistent representations of activities or other occurrents," it is important to recognize that one "can expect to find representations at many different levels" [1]. These are not just levels in the functional hierarchy of records but also levels of representation. Digital records can be considered and encountered at levels ranging from aggregations of records down to bits as physically inscribed on a storage medium; each level of representation can provide distinct contributions to the information and evidential value of records [2]. There is a substantial body of information within the underlying data structures of computer systems that can often be discovered or recovered, revealing new types of records or essential metadata associated with existing record types. The multiple representation levels of digital materials have significant implications for all archival functions [3].

The second primary consideration is the direct "machine-readability" of digital materials. This allows archivists to use software to identify, extract and manipulate patterns in and between records in ways that would not be feasible with analog records. Appraisal "is an iterative [process] that becomes progressively more refined as more information about the records and their context of creation becomes available" [29]. Contextual information takes a variety of forms, and it can be challenging to identify and represent [4]. Because contextual information—both embedded in digital objects and in relationships between them—can be detected and captured using software, there is potential to better inform and facilitate archival practices. However, such activities require specific machine instructions. Following "requirements-based workflows necessitates appraisal iterations that are more controlled than those for analogue records" [29].

In 2015, there was a breakout discussion about appraisal at an event called Capture Lab. Two major themes from this discussion were: (1) there are numerous data elements within born-digital materials that could be used (but currently are not used) to support more effective and efficient appraisal processes, and (2) appraisal is not a specific point in a digital curation workflow but is instead something that happens at numerous points throughout the process. This suggests that, rather than trying to develop one, monolithic system devoted specifically to selection and appraisal, the goal should instead be to incorporate these tools and methods into environments where library, archives and museum (LAM) professionals are carrying out their workflows more generally.

## II. PREVIOUS WORK

As the abundance and diversity of documentation has grown, so has the importance of selection and appraisal. The explosion of record volume in the 1930s and 1940s served as a catalyst for a professional literature on archival appraisal and the formation of records management as a distinct field of endeavor. Beginning in the 1970s, this discussion turned to electronic records (or what were then called machine-readable records) [5][6], and the literature on the appraisal of electronic records has continued to grow slowly since then. Authors have elaborated a variety of criteria and principles to consider when engaged in appraisal, but relatively few have investigated the use of software to support such decision-making.

One of the first efforts to apply software to archival appraisal was a study by Gillilan, in which she elicited knowledge from domain experts and then attempted to develop an expert system [7][8]. She was unable to identify a consensus on appraisal rules or principles. This suggests that software to support appraisal should allow archivists to make individual

decisions based on iterative feedback, rather than attempting to replace the human decision-maker with software. Software for selection and appraisal can take the form of targeted tools to support specific assessments or decisions, rather than necessarily being full-fledged decision-support systems. For example, the Wellcome Library has investigated the use of common tools to facilitate aspects of appraisal: DROID to identify file types that can be discarded and md5 hashes for deduplication [28].

Lee gave a conference presentation in 2000 on the topic of computer-assisted approach, in which he elaborated various types of software that archivists could use [9]. However, he did not follow up with any testing or implementation at that time. Similarly, in 2010, Harvey and Thompson discussed the prospects of using software for appraisal and re-appraisal, but they did not offer any follow-up implementation of these ideas [10].

The area in which records professionals have most thoroughly translated appraisal criteria into specific software actions is web archiving [32], which requires machine-actionable instructions. Archivists must negotiate a variety of "crawl modalities" [30]. Four fundamental parameters to define are: environments crawled, access points from those environments used as crawling or selection criteria, threshold values for scoping capture within given access points, and frequency of crawls [31]. In 2005, Pearce-Moses and Kaczmarek developed the "Arizona Model" that was based on mapping records retention schedule series to web sites [11]. While the Arizona Model has not been widely adopted, the idea of applying retention schedule criteria to web crawling has carried on. The State Archives of North Carolina developed a set of guidance documents for mapping records retention categories to specific web archiving actions [12].

In light of recent case law validating its use, there has also been investigation into the use of "predictive coding" (text classification based on natural language processing (NLP) rather than simple string matches) to identify subsets of records that warrant attention [13]. The National Archives of the UK has carried out an exploratory investigation of such approaches [14]. Several state governments in Australia have conducted similar investigations.

Other authors have reported on email appraisal efforts. Much of this work has involved the use of software in some aspects of the workflow but then application of manual process for the selection of messages. One example is the Library of Virginia's processing of email from Governor Tim Kaine's administration [15]. Cocciolo also conducted a case study that involved manual application of an email selection rubric [16].

A selected set of projects have investigated the use of software to select email. Vinh-Doyle reports on exploratory efforts to use EnCase, a commercial digital forensics software suite, to identify email of continuing value [17]. This work identified some interesting factors for consideration but did not establish a set of methods or tools for use by other institutions. The Illinois State Archives, in partnership with the University of Illinois and with funding from the National Historical Publications and Records Commission (NHPRC), has attempted to use predictive coding to identify and provide appropriate access to the email messages of state agencies, based on the National Archives and Records Administration's Captstone Email approach.[1] Vellino and Alberts analyzed appraisal behaviors of eight records management experts to train a series of classifiers to identify email messages with business value; they found that the dominant discriminating factors to be textual features from the e-mail body and subject field (as opposed to values in the rest of the email header) [27].

## III. OPPORTUNITIES

There are numerous opportunities for further advancing computer-assisted appraisal and selection.

### *A. Digital Forensics*

The application of digital forensics methods in LAMs has been advanced by several projects funded by the Andrew W. Mellon Foundation. These included the Computer Forensics and Born-Digital Content in Cultural Heritage Collections project [18]; the BitCurator project, which packaged an open source software environment allowing users to apply digital forensics methods to collections [19][20]; BitCurator Access, which developed tools to assist LAMs in both redacting and providing access to data from disk images [21]; and BitCurator NLP, which has developed NLP-supported tools to identify and report on entities of interest within born-digital collections.[2]

Dates and chronological relationships can play a vital role in appraisal decisions. Over the past decade, open source digital forensics tools designed to support timeline analysis have gained significant traction. Today, tools such as log2timeline[3] are an important resource for forensic investigators due to their ability to collate disparate and inconsistently formatted metadata from many different sources and organize it to support typical activities within their workflow. The rationale is simple: "Arranging events chronologically is a good way of telling a clear, concise story" [22]. As powerful as these tools are, however, their implementations focus on organizing this metadata in a format that allows a forensic investigator to tag otherwise intractable volumes of material quickly. They are not intended to support access mechanisms or provide more holistic views of the lifecycle of the materials. However, archivists often are interested in a more holistic view, as their work does not terminate with the successful prosecution of a case, but may support the work of researchers interested in building a map of connections within the materials. There are substantial opportunities to improve metadata export and timelining facilities for collections containing born-digital records, as timestamps often are automatically recorded (e.g. in email headers, filesystem attributes of files) during their production and use.

---

[1] https://www.uillinois.edu/cio/services/rims/about_rims/projects/processing_capstone_email_using_ predictive_coding/

[2] https://github.com/BitCurator/bitcurator-nlp

[3] https://github.com/log2timeline/plaso/

### B. Natural Language Processing

One of the primary motivations for applying digital forensics tools in archives is to capture and provide access to contextual information. For example, the original filesystem attributes associated with files (e.g. directory paths, timestamps) can be essential to understanding their provenance and original order. However, there are many other types of contextual information that can be vital to making sense and meaningful use of digital objects. These include nine classes of contextual entities: object, agent, occurrence, purpose, time, place, form of expression, concept/abstraction and relationship [4]. In a study of reference questions submitted to archives, Duff and Johnson found that most information requests were based on "proper names, dates, places, subject, form, and, occasionally, events when composing their information request" [23]. In their study of genealogists, Duff and Johnson identified information seeking practices that were focused primarily on names, places and time periods [24]. If appraisal decisions are to be informed by (among other factors) the components of records that will be relevant to users, then application of NLP to identify entities related to records could be beneficial.

There are many mature open source natural language processing platforms that provide web services and RESTful application programming interfaces (APIs) and integration with industry-standard testing and training corpora. Production-quality open source software toolkits for natural language processing include OpenNLP (Java-based) and NLTK, Pattern, and spaCy (Python-based). Some of these platforms have been used in projects specifically targeted at LAMs, but the use cases are often quite specific.

One such project has been ePADD. The ePADD software supports processing collections of email by using a customized Named Entity Recognition engine to identify correspondents within email. A web publication from the group notes: "Not satisfied with other open source NER engines, including the Stanford NER and Apache OpenNLP, the ePADD development team created their own engine … to help identify and disambiguate correspondents within the corpus … [and] ensures persons that occur within the email archive who are also correspondents are weighted more heavily in this ranking." At the time of writing it appears that this engine is integrated directly into the ePADD application, rather than as a reusable library.

In the digital humanities, there have been many years of work on applying NLP to the content of primary sources. Projects in the field often focus on specific areas of NLP, such as named entity recognition (NER) and topic modeling to provide researchers with meaningful views of the people, organizations, and events described within a formal collection or data gathered from the Web. There is great potential to apply these methods more widely to archival collections, in order to identify and expose the sorts of contextual entities discussed above.

### C. Machine Learning

Machine learning (ML) can allow software to progressively improve its performance on given tasks without the improvements being explicitly programmed; the software learns by building and refining a statistical model based on training data. Archives often have large and diverse collections and limited human resources, and such approaches could benefit processing workflows by reducing the time required to triage materials and automating certain classification tasks. One of the challenges is that generating training data can be very labor-intensive. In order for an ML model to classify a digital object as a particular type of record (e.g., official correspondence within a specific records series) or a non-record (anything outside the scope of preservation), tens of thousands of records correctly annotated by a human archivist might be required for training.

Active learning (AL) is a process in which the software tries to prioritize instances for human review that are most likely to inform the underlying model. While this can improve performance, it still requires either a large amount of training data or a significant number of human expert judgements. One study has demonstrated a "novel interactive learning algorithm that is capable of directly acquiring domain knowledge from human experts by allowing them to articulate the evidence that leads to their sense tagging decisions (e.g., the presence of indicative words in the context that suggest the sense of the word)" [25]. This knowledge is then applied in subsequent learning processes to help the algorithm achieve desirable performance with fewer iterations." While they applied this approach specifically to word sense disambiguation in medical records, such interactive machine learning based on multiple forms of human input holds great promise for archival appraisal.

Another promising domain for machine learning vital to selection and apprisal is review for sensitivity. Electronic records often contain personal identifiers, discussions of sensitive subjects, or other information that may be subject to restriction or redaction. The Presidential Electronic Records Pilot System (PERPOS) Project has investigated several technical approaches (e.g. automatic identification of speech acts) to support assignment of access restrictions and declassification [26]. In October 2018, the National Library of Scotland, with support from Arts and Humanities Research Council (AHRC) though the Scottish Graduate School for Arts and Humanities, Information Studies at the University of Glasgow, began a study to "use innovative methods for handling sensitive information, focusing on compliance with legal obligations (e.g. data protection)" and "investigate broader concerns, such as cost and data ethics of incorporating AI in data handling."[4]

[4] https://www.gla.ac.uk/colleges/arts/graduateschool/fundingopportunities/aistudentship

## IV. CONCLUSION

Selection and appraisal are vital functions of archives, but there has been relatively little attention focused on computational methods to support those functions. Appraisal involves human judgements based on a variety of social, institutional and technical factors. Replacing such judgements with software is neither desirable nor realistic. However, enhancing and better supporting selection and appraisal is a goal worthy of further research and development.

## REFERENCES


[1] G. Yeo, “Concepts of record (2): prototypes and boundary objects,” American Archivist, vol. 71, no. 1, pp. 118–143, 2008.

[2] C. Lee, “Digital curation as communication mediation,” in Handbook of Technical Communication, A. Mehler, L Romary, and D. Gibbon, Eds. Berlin, Mouton De Gruyter, 2012, pp. 507–530.

[3] C. Lee, "Archival application of digital forensics methods for authenticity, description and access provision," Comma, vol. 2, no. 14, pp. 133-139, 2012.

[4] C. Lee, "A Framework for Contextual Information in Digital Collections," Journal of Documentation, vol. 67, no. 1, pp. 95-143, 2011.

[5] C. Dollar, "Appraising Machine-Readable Records," American Archivist, vol. 41, no. 4, pp. 423-430, 1978.

[6] H. Naugler, The Archival Appraisal of Machine- Readable Records: A RAMP Study with Guidelines. Paris: General Information Programme and UNISIST United Nations Educational Scientific and Cultural Organization, 1984.

[7] A. Gilliland-Swetland, "Development of an Expert Assistant for Archival Appraisal of Electronic Communications: An Exploratory Study," PhD Dissertation, University of Michigan, 1995.

[8] A. Gilliland, "Designing Expert Systems for Archival Evaluation and Processing of Computer-Mediated Communications," in Research in the Archival Multiverse, A. Gilliland, S. McKemmish and A. Lau, Eds. Clayton, Australia: Monash University Publishing, 2016, pp. 686-722.

[9] C. Lee, "Computer-Assisted Appraisal of Electronic Records," joint meeting of Midwest Archives Conference and Mid-Atlantic Region Archives Conference, Cleveland, OH, October 19, 2000.

[10] R. Harvey and D. Thompson, "Automating the Appraisal of Digital Materials," Library Hi Tech, vol. 28, no. 2, pp. 313-322, 2010.

[11] R. Pearce-Moses and J. Kaczmarek. "An Arizona Model for Preservation and Access of Web Documents," DttP, vol. 33, no. 1, pp. 17-24, Spring 2005.

[12] North Carolina Department of Cultural Resources, ”Standard for Automated Web Site Capture,” July 17, 2006, https://files.nc.gov/dncr-archives/documents/files/websitestandards.pdf.

[13] D. Smith, ”Thinking Outside the Box: Use of Predictive Coding as a RIM Tool,” Information Management, vol. 47, no. 1, pp. 30-32, 46, 2013.

[14] "The Application of Technology-Assisted Review to Born-Digital Records Transfer, Inquiries and Beyond." National Archives of the UK, 2016.

[15] R. Christman and S. Page, "Addressing the challenge of the governor's e-mail," MAC Newsletter, vol. 43, no. 1, 2010.

[16] A. Cocciolo, "Email as cultural heritage resource: appraisal solutions from an art museum context," Records Management Journal, vol. 26, no. 1, pp. 68-82, 2016.

[17] W. Vinh-Doyle, “Appraising email (using digital forensics): techniques and challenges," Archives and Manuscripts, vol. 45, no. 1, pp. 18-30, 2017.

[18] M. Kirschenbaum, R. Ovenden, and G. Redwine, Digital Forensics and Born-Digital Content in Cultural Heritage Collections. Washington, DC: Council on Library and Information Resources, 2010.

[19] C. Lee, K. Woods, M. Kirschenbaum, and A. Chassanoff, "From Bitstreams to Heritage: Putting Digital Forensics into Practice in Collecting Institutions," September 30, 2013.

[20] C. Lee, P. Olsen, A. Chassanoff, K. Woods, M. Kirschenbaum, and S. Misra, “From code to community: building and sustaining BitCurator through community engagement,” September 30, 2014.

[21] K. Woods, C. Lee, O. Stobbe, T. Liebetraut, and K. Rechert, “Functional access to forensic disk images in a web service,” in Proceedings of the 12th International Conference on Digital Preservation. Chapel Hill, NC: University of North Carolina, School of Information and Library Science, 2015, pp. 191-195.

[22] D. Edwards, “Computer Forensic Timeline Analysis with Tapestry,” SANS Institute Reading Room, November 12, 2011.

[23] W. Duff and C. Johnson, “A virtual expression of need: an analysis of e-mail reference questions,” American Archivist, vol. 64, pp. 43-60, 2001.

[24] W. Duff and C. Johnson, “Where is the list with all the names? information-seeking behavior of genealogists,” American Archivist, vol. 66, pp. 79-95, 2003.

[25] Y. Wang, K. Zheng, H. Xu, and Q. Mei, “Interactive medical word sense disambiguation through informated learning,” Journal of the American Medical Informatics Association, vol. 25, no. 7, pp. 800-808, 2018.

[26] W. Underwood, M. Hayslett, S. Isbell, S. Laib, S. Sherrill, and M. Underwood, "Advanced decision support for archival processing of presidential electronic records: final scientific and technical report," Georgia Institute of Technology, 2009.

[27] A. Vellino and I. Alberts, "Assisting the Appraisal of E-Mail Records with Automatic Classification," Records Management Journal, vol. 26, no. 3, pp. 293-313, 2016.

[28] V. Sloyan, "Born-digital archives at the Wellcome Library: appraisal and sensitivity review of two hard drives," Archives and Records, vol. 37, no. 1, pp. 20-36, 2016.

[29] C. Mumma, G. Dingwall, and S. Bigelow, "A first look at the acquisition and appraisal of the 2010 olympic and paralympic winter games fonds: or, select * from Vanoc_Records as archives where value=’true’;," Archivaria, vol. 72, pp. 93-122, 2011.

[30] E. Summers and R. Punzalan, “Bots, Seeds and People: Web Archives as Infrastructure,” Proceedings of the 20th ACM Conference on Computer Supported Collaborative Work, Portland, Oregon: ACM, 2017.

[31] C. Lee, “Collecting the Externalized Me: Appraisal of Materials in the Social Web,” in I, Digital: Personal Collections in the Digital Era, C. Lee, Ed. Chicago, Society of American Archivists, 2011.

[32] C. Post, "Building a living, breathing archive: a review of appraisal theories and approaches for web archives," Preservation, Digital Technology and Culture, vol. 46, no. 2, pp. 69-77, 2017.