Processing Capstone Email Using Predictive Coding

Brent West
University of Illinois
506 S Wright St, M/C 359
Urbana, IL 61801 USA
+1 217-265-9190
bmwest@uillinois.edu

Joanne Kaczmarek
University of Illinois
506 S Wright St, M/C 359
Urbana, IL 61801 USA
+1 217-333-6834
jkaczmar@illinois.edu

ABSTRACT

Email provides a rich history of an organization yet poses unique challenges to archivists. It is difficult to acquire and process, due to sensitive contents and diverse topics and formats, which inhibits access and research. We plan to leverage predictive coding used by the legal community to identify and prioritize sensitive content for review and redaction while generating descriptive metadata of themes and trends. This will empower records creators, archivists, and researchers to better understand, synthesize, protect, and preserve email collections. Early findings and information on collaborative efforts are shared.

Keywords

Archives; Continuous active learning; Dataless classification; Descriptive metadata; E-discovery; FOIA; Metrics-based reappraisal; MPLP; Natural language processing; Restricted records; Self-appraisal; Sustainable digital preservation; Technology-assisted review.

1. INTRODUCTION

The Records and Information Management Services (RIMS) office of the University of Illinois is leading a project to help archivists preserve email messages of enduring value, beginning with those of the University's senior administrators [1]. Email messages of senior administrators are the modern equivalent of correspondence files, long held to have enduring value for administrators and researchers alike. However, email presents unique accessioning challenges due to its quantity, file formats, conversation threads, inconsistent filing, links and attachments, mix of personal and official communications, and exposure of sensitive content.

The quantity and mix of content, as well as the inability to rely upon administrators to consistently identify messages of enduring value, led RIMS to explore the Capstone approach developed by the United States National Archives and Records Administration (NARA) [2] to stem the loss of significant correspondence. The Capstone approach offers an option for agencies to capture most of the email from the accounts of officials at or near the head of an agency without detailed consideration of the content.

Although this approach can help to ensure that significant correspondence is retained, Capstone is just the first step in the overall curation lifecycle [3] at a scale which necessitates More Product, Less Process [4]. Processes and tools such as Preservica exist to acquire, ingest, store, transform, and even provide access to email. However, unmet lifecycle challenges of email include the identification of restricted records as a prerequisite to public access and the reappraisal of non-archival messages in heterogeneous email collections. Techniques such as Metrics-Based Reappraisal [5] can sustainably inform reappraisal decisions for a variety of digital collections. However, we propose a new methodology to address both unmet challenges.

2. PREDICTIVE CODING

2.1 E-discovery

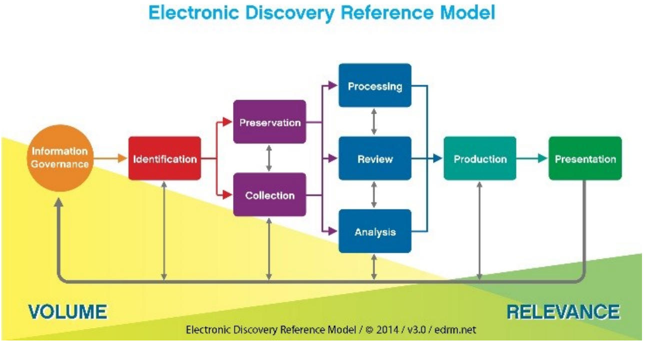

Figure 1. Electronic discovery reference model. [6]

Electronic discovery is a "process in which electronic data is sought, located, secured, and searched with the intent of using it as evidence in a civil or criminal legal case" [7]. Courts require good faith efforts to discover and produce relevant evidence for the opposing party to a lawsuit. E-discovery provides attorneys insight into both their case and their opponents' case, uncovering critical evidence that can resolve the case in one's favor. With potentially millions of dollars on the line, the legal community has a substantial incentive to conduct a thorough review. At the same time, courts recognize that the burden of discovery must be proportional to the potential evidentiary value, the amount in dispute, and the resources of the parties. Even so, e-discovery is expensive with mean costs comprising 73% or $22,480 per gigabyte reviewed [8]. To combat these high costs and provide a competitive advantage, attorneys and courts are increasingly turning to technology to make the review process more efficient.

2.2 Technology-Assisted Review

Technology-assisted review (TAR) enhances the heretofore manual review of potentially relevant records by providing insight into data collections. TAR allows attorneys to more quickly locate potentially responsive documents and cull that list based on various attributes to narrow and prioritize review and redaction efforts. TAR tools often feature de-duplication, email threading, full-text search of messages and common attachments, and pattern and trend visualizations. Increasingly, TAR tools are providing natural language processing and machine learning features to cluster documents by topics and identify hidden relationships through predictive coding.

2.3 Predictive Coding

Predictive coding leverages artificial intelligence algorithms to locate relevant documents. Relevant documents that have been assessed manually by humans are processed by the algorithms to automatically assess the relevance of other documents in a large collection. In an iterative process of automated assessment and review, the software begins to learn what attributes make a document relevant, increasing the capacity to quickly identify documents of most interest. A variation of this approach is known as Continuous Active Learning [9] where the process is repeated until no further items shown are relevant. This ability to automatically categorize hundreds of thousands to millions of documents greatly enhances the effectiveness of document review, allowing attorneys to prioritize their review around the most valuable or sensitive content.

In a sense, predictive coding is automating the generation of topical descriptive metadata. The identification of documents that are relevant to particular topics allows archivists to prioritize the review of a large email collection and identify restricted records and non-archival items. For instance, items related to personnel matters or family medical leave could be redacted, restricted, or purged as appropriate. At the same time, categorized messages would be of immense value to researchers who would no longer have to be as concerned that relevant messages were overlooked in a manual or keyword search.

3. WORKFLOW

3.1 Capstone

The University Archivists have identified approximately 0.1% of its employees as senior administrators for whom most or all email should be retained for its institution-wide value. Another 1% have been identified as mid-level administrators that will frequently have correspondence of significant value to their area of responsibility but do not necessitate retention in bulk.

It is critical to the Capstone approach to inform relevant email account owners of the approach and of the historical value of their correspondence. This opportunity should also be used to address any concerns about the appraisal, transfer, or access restriction processes as well as establish a recurring schedule for ingests. Owners will benefit from specific guidance about items of archival value as well as general email management best practices.

3.2 Transfer

Email transfers frequently occur upon retirement or separation of the individual, which is often when records are most at risk of loss. At a minimum, the office and the successor should retain a copy of important records and correspondence for business continuity purposes.

After a clearly defined period, perhaps 3-6 years after separation, the email should be transferred to the custody of the archives. In a Microsoft Exchange environment, this may be accomplished in a native PST format, possibly using an external hard drive. If possible, custodians should include information describing the main categories of subjects that exist within the correspondence as well as any forms of confidential information that may exist. Custodians may choose to pre-screen the content in order to withhold active or sensitive topics until a later date.

3.3 Processing

3.3.1 Identify

Topics of interest should be identified from transferred email collections. This may be developed through traditional record series and folder lists, sampling informed by the originating office, or using techniques such as data-less classification [10] to gain insights into unknown datasets. De-duplication of identical or nearly identical messages (e.g., sender vs. recipient copy) is also useful at this stage.

3.3.2 Describe

Using a predictive coding tool such as Microsoft's Advanced eDiscovery for Office 365 (formerly Equivio), the messages will be associated with the topics identified above through an iterative training process. Although results may be available through a quick review of as few as 1,000 messages, a greater set of training data will produce more reliable results. Feedback provided during the training process will help determine when training is complete. It is important to note that text must be extracted from the attachments to successfully categorize the document.

3.3.3 Redact

A prioritized review may now be conducted to focus efforts on likely candidates for confidential information. For instance, attorney-client communications and student advising records should be reviewed more carefully while press releases and mass mailings likely require less stringent review. Tools such as Identity Finder or Bulk Extractor may help locate regular forms of personally identifiable information. In addition, review-on-demand services could be offered to provide quick access to researchers while ensuring that access to confidential information is restricted.

3.3.4 Preserve

Multiple tools exist to preserve email, an especially important function given the proprietary and sometimes volatile nature of PST files. Preservica, for instance, uses Emailchemy to extract messages and attachments from PST files and convert the messages to the plain-text EML format. Preservica also supports multiple manifestations, allowing redacted versions of documents in popular formats for public access and un-redacted versions in native and sustainable formats for preservation.

3.4 Access

Although Preservica could also be used to provide online access through its Universal Access feature, many institutions may prefer to maintain offline access using a terminal in the archives. A hybrid of this might utilize the redacted view feature of ePADD [11] to provide limited online keyword search capabilities and general trend visualizations without exposing the full content of a message. Full access may be facilitated in a native email client at the archives terminal. A confidentiality agreement could also be used to further protect against the disclosure of overlooked restricted content.

4. NEXT STEPS

The long-term preservation of digital content presents many challenges to the archival community. The continued custodial responsibilities needed to ensure that content is preserved over time and remains reliably accessible will require thoughtful decisions to be made regarding what content to prioritize. If successful, the use of predictive coding to process Capstone email may provide administrators, researchers, and archivists alike with tools that can assist in making more informed decisions using active and inactive content, responding more swiftly and accurately to requests under freedom of information laws, and performing a limited self-appraisal to identify messages that are of a personal nature or that warrant access restrictions.

During the summer and fall of 2016, the University of Illinois is collaborating with the Illinois State Archives to manually categorize a subset of topics for a 2 million message collection from former Illinois gubernatorial administrations. The results of this effort will be used as part of a National Historical Publications & Records Commission-funded project to evaluate the effectiveness of various predictive coding tools to supplement traditional digital archival methods and ultimately to accession, describe, preserve, and provide access to state government electronic records of enduring value.

5. ACKNOWLEDGMENTS

The authors would like to thank Dave Joens, Kris Stenson, Maura Grossman, Gordon Cormack, Dan Roth, Jordan Phoenix, Shashank Gupta, Roger Christman, NexLP, and the University of Illinois Office of University Relations for their support and collaboration to provide a sustainable path for the digital preservation of email correspondence of enduring value.

6. REFERENCES

[1] University of Illinois Records and Information Management Services. 2015. Preserving Email Messages of Enduring Value. http://go.uillinois.edu/capstone.

[2] U.S. National Archives and Records Administration. 2013. NARA Bulletin 2013-02. http://www.archivess.gov/records-mgmt/bulletins/2013/2013-02.html.

[3] Digital Curation Centre. 2008. DCC Curation Lifecycle Model. http://www.dcc.ac.uk/resources/curation-lifecycle-model.

[4] Greene, M. A. and Meissner, D. 2005. More Product, Less Process: Revamping Traditional Archival Processing. In *The American Archivist,* 68 (Fall/Winter 2005), 208–263. http://www.archivists.org/prof-education/pre-readings/IMPLP/AA68.2.MeissnerGreene.pdf.

[5] University of Illinois Records and Information Management Services. 2014. Metrics Based Reappraisal. http://go.uillinois.edu/rimsMBR.

[6] EDRM. 2014. EDRM Stages. http://www.edrm.net/resources/edrm-stages-explained.

[7] TechTarget. 2010. Electronic discovery (e-discovery or ediscovery). http://searchfinancialsecurity.techtarget.com/definition/electronic-discovery.

[8] RAND Institute for Civil Justice. 2012. Where the Money Goes. http://www.rand.org/content/dam/rand/pubs/monographs/2012/RAND_MG1208.pdf.

[9] Grossman, M. R. and Cormack, G. V. 2016. Continuous Active Learning for TAR. In *Practical Law* (April/May 2016), 32-37. http://cormack.uwaterloo.ca/cormack/caldemo/AprMay16_EdiscoveryBulletin.pdf.

[10] University of Illinois Cognitive Computation Group. 2014. Dataless Classification. https://cogcomp.cs.illinois.edu/page/project_view/6.

[11] Stanford University. 2015. ePADD. https://library.stanford.edu/projects/epadd.