Email Preservation at Scale: Preliminary Findings Supporting the Use of Predictive Coding

Joanne Kaczmarek
University of Illinois
506 S Wright St Rm 450, M/C 359
Urbana, IL 61801 USA
+1 217-333-6834
jkaczmar@illinois.edu

Brent West
University of Illinois
506 S Wright St Rm 450, M/C 359
Urbana, IL 61801 USA
+1 217-265-9190
bmwest@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 content and diverse topics and formats, which inhibits access and research. Predictive coding alleviates these challenges by using supervised machine learning to: augment appraisal decisions, identify and prioritize sensitive content for review and redaction, and generate descriptive metadata of themes and trends. Following the authors' previous work which describes the project at its inception, preliminary findings support the use of predictive coding as an effective tool to enable digital preservation at scale. Specific tools, methodologies, and human factors that affect their success are discussed.

Keywords

Access; Active learning; Appraisal; Automatic identification; Big data; Case studies and best practice; Data identification; Descriptive metadata; Digital archives; Digital humanities; Digital preservation; E-discovery; Email archiving; File formats; Ingest; Lessons learned; Natural language processing; PII; Preservation strategies and workflows; Privacy; Redaction; Restricted records; Scalability; Sustainability; Self-appraisal; Software-as-a-service; Supervised learning; Technological infrastructure; Technology-assisted review; Text mining; Unstructured data.

1. INTRODUCTION

The Records and Information Management Services (RIMS) office of the University of Illinois, in conjunction with the University Library, is working with the Illinois State Archives (ISA) on a project to acquire, process, and provide access to a collection of email messages from senior government officials of the State of Illinois [1]. The project is generously funded by a three year grant through the National Historical Publications and Records Commission (NHPRC). A unique aspect of this project is the evaluation of commercial tools which may help the archives community effectively process large collections of email messages and other similarly diverse corpuses. In particular, the project will discover how tools developed by the legal community for electronic discovery (e-discovery) might augment the preliminary archival review and increase processing output. Doing so will assist archivists in making appraisal and preservation decisions for large heterogeneous email collections with greater confidence and precision to the item level. This empowers records creators, archivists, and researchers to better understand, synthesize, protect, and preserve email collections.

1.1 Challenges

For this project, e-discovery tools are being assessed using email messages secured through the implementation of the Capstone approach developed by the United States National Archives and Records Administration (NARA) [2]. The Director of the ISA considers email messages from senior administrators in key state agency offices to be the modern equivalent of subject or general correspondence files, long held to have enduring value for administrators and researchers alike. However, email continues to present unique accessioning challenges due to a variety of factors, including its: large volume and duplication; conversation threads; diverse file formats and attachments; links to external documents and resources; inconsistent classification and mixing of personal, informal, and official communications; and the prevalence of sensitive content. Because of these challenges, email messages remain absent from the State's archival holdings which poses a substantial risk of loss.

The Capstone approach is just the first step in ensuring that significant correspondence is retained. It offers an option for the ISA to preserve most of the email from the accounts of officials at or near the head of an agency without detailed consideration of the content. But without such consideration, much of the content may never be made publicly available. The sheer volume of email collections will require automated review processes if content is to be made available on an ongoing basis. This project, for example, consists of over 500 GB and 5M messages from 68 individuals from 2003-2010 with another 113 accounts requested but not found. Similar results have been observed on a related Capstone project for senior administrators at the University of Illinois [3]. In addition to concerns about volume, review processes must also reliably identify various types of sensitive data, both known and unknown, as well as uniquely identify each item in such a way as to maintain relationships to the data source (e.g., a PST file) and its derivative preservation and access formats (e.g., EML) and systems.

1.2 Predictive Coding

The cultural heritage community has made significant headway in recent years in the automated review and batch processing of unstructured data such as email. Workflows and open source tools have been developed through projects such as ePADD [4] and TOMES [5]. Commercial tools such as Preservica provide a means of ingesting, storing, transforming and even accessing email messages. Nonetheless, challenges remain when scaling efforts across large collections and diverse domains. Identifying various types of restrictions that must be placed on email collections prior to making them publicly available is a notable challenge at scale as is how to easily reduce the number of non-archival messages. This project addresses these long-standing challenges using tools that employ some level of predictive coding or active learning techniques previously described by the authors in their 2016 iPRES paper [6].

Predictive coding uses computer-generated statistical models to locate documents relevant to a particular inquiry based on a manual human review of a sample of documents from the collection. From this coded sample, algorithms are trained and a relevance score is calculated for each document in the collection. Through an iterative process of coding additional samples, scoring the documents, and comparing the calculated scores to the human coding for a control group, the software begins to learn what attributes make a document relevant. Through this process an assessment of the performance of the model can be made to increase the capacity to quickly identify documents of most interest. These tools can also be used to automate the generation of descriptive metadata through concept clustering thereby helping identify messages that should remain restricted from access. This ability to automatically categorize hundreds of thousands to millions of documents may prove to reliably reduce the time needed by the archival community to identify, preserve, and provide appropriate access to archival email messages.

2. TOOLS REVIEW

2.1 Identification

The project team built a template (see Fig. 1) for recording information about tools modeled after a similar template developed by the Digital POWRR project [7]. Features desired or required for an email analysis workflow were identified for three high-level categories: pre-processing, content analysis, and content preparation. Nineteen tools, both open source and commercially available, were identified by reviewing rankings of products offered through Gartner's Magic Quadrant [8] and through consultation with members of the archival community. Some tools offer very specific functionality while others offer more of a suite of services and features. The review process consisted of searching for content provided by user groups, subject matter experts, documentation available via tool websites, and demonstrations. Based on the project focus of predictive coding, the project selected four commercial tools for hands-on analysis and two open source tools. Commercial tools include Microsoft's Office 365 Advanced eDiscovery [9], OpenText's Recommind [10], FTI Consulting's Ringtail [11], and Luminoso Analytics [12], and the open source tools are ePADD and the TAR Evaluation Toolkit [13].

2.2 Assessment

After acquiring licenses for each, the project team loaded a subset of email into Advanced eDiscovery, Recommind, Ringtail, and finally Luminoso. Exploring the predictive coding capabilities of each tool has been prioritized, but each tool has a great amount of functionality beyond predictive coding that may be of interest to the archival community.

2.2.1 *Advanced eDiscovery*

Due to preexisting Microsoft licensing at the University of Illinois, the team had access to Advanced eDiscovery at a very low cost (the full retail cost is $420 per year [14]) at the start of the project and so began the initial tool assessments with it. A small PST file (1.1GB) was uploaded into a separate Office 365 tenant for the project. Preliminary focus of the assessment was to learn how the tool labels email messages based on "themes" automatically identified through analysis of the content found in the body of the messages. Once themes were identified, the team selected several themes and began to assess the dataset by tagging for these themes. With this particular dataset, we found too few relevant messages for each theme selected to be able to create a useful training set. A review to identify documents in a broader theme such as "restricted" was also unsuccessful in initial testing. We will be revisiting Advanced eDiscovery with a larger dataset and with different approaches in the future.

2.2.2 *Recommind*

Header analysis as well as content analysis is an option for email threads in Recommind (see Fig. 2) and many options exist for how documents are displayed and reviewed. Icons provide information about whether a document has parent or child relationships, whether it is part of a thread, or whether it has duplicates or near duplicates. Highlighting is available to bring quick attention to specific terms such as those associated with search terms, training terms, or concept terms. Various dashboards are also available to assess the productivity of document review by various attributes (see Fig. 3). The predictive coding dashboard assesses progress towards locating all desired documents for an inquiry (see Fig. 4).

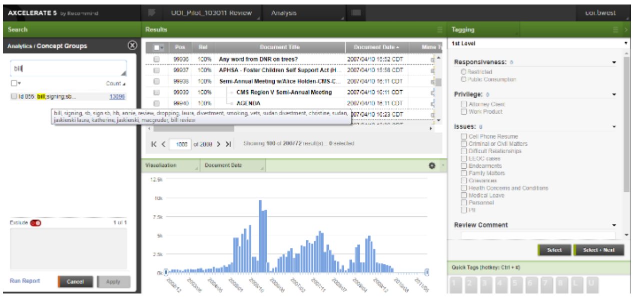
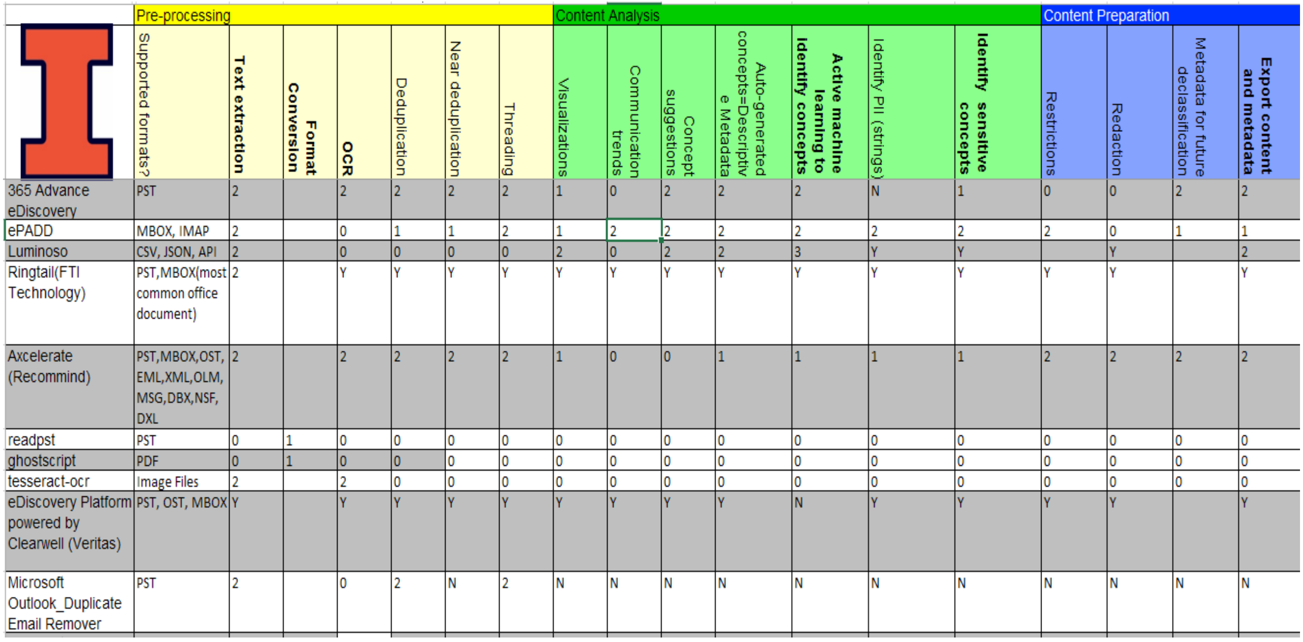


Figure 2. Recommind analysis, review, and concepts.



	Pre-processing							Content Analysis							Content Preparation			
	Supported formats?	**Text extraction**	**Format Conversion**	**OCR**	Deduplication	Near deduplication	Threading	Visualizations	Communication trends	Concept suggestions	Auto-generated concepts=Descriptive Metadata	**Active machine learning to identify concepts**	Identify PII (strings)	**Identify sensitive concepts**	Restrictions	Redaction	Metadata for future declassification	**Export content and metadata**
365 Advance eDiscovery	PST	2		2	2	2	2	1	0	2	2	2	N	1	0	0	2	2
ePADD	MBOX, IMAP	2		0	1	1	2	1	2	2	2	2	2	2	2	0	1	1
Luminoso	CSV, JSON, API	2		0	0	0	0	2	0	2	2	3	Y	Y		Y		2
Ringtail(FTI Technology)	PST,MBOX(most common office document)	2		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y		Y
Axcelerate (Recommind)	PST,MBOX,OST, EML,XML,OLM, MSG,DBX,NSF, DXL	2		2	2	2	2	1	0	0	1	1	1	1	2	2	2	2
readpst	PST	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ghostscript	PDF	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
tesseract-ocr	Image Files	2		2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
eDiscovery Platform powered by Clearwell (Veritas)	PST, OST, MBOX	Y		Y	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	Y	Y		Y
Microsoft Outlook_Duplicate Email Remover	PST	2		0	2	N	2	N	N	N	N	N	N	N	N	N	N	N

Figure 1. Capstone project tools grid. [15]

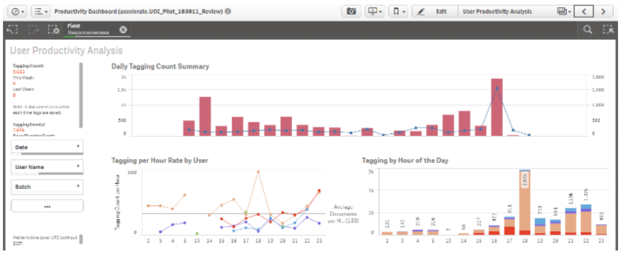
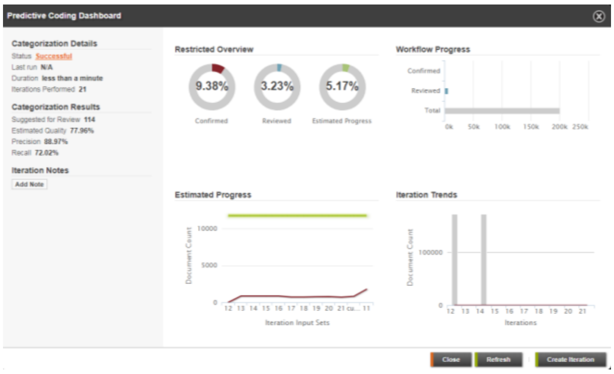

User Productivity Analysis
Daily Tagging Count Summary
Tagging per Hour Rate by User
Tagging by Hour of the Day
Date
User Name
Batch

Figure 3. Recommind productivity dashboard.



Categorization Details
Status Successful
Last run N/A
Duration less than a minute
Iterations Performed 21

Categorization Results
Suggested for Review 114
Estimated Quality 77.96%
Precision 88.97%
Recall 72.02%

Iteration Notes

Figure 4. Recommind predictive coding dashboard.

2.2.3 *Ringtail*

Ringtail has many similar features to Recommind. In addition to a variety of ways in which multi-faceted searches can be performed based on an extensive list of extracted attributes, Ringtail provides "mines" and "cubes" as a way to explore concept clusters and documents with similar attributes. Ingestion workflows provide an extensive array of options and audit logs. The predictive models in Ringtail provide a high level of insight into the performance of a model, allowing one to fine tune the desired level of recall, precision, and accuracy based on the manual review effort required and tolerance for mistakes (see Fig. 5 and Fig. 6). Ringtail has fine-grained security capabilities and extensive training resources available.

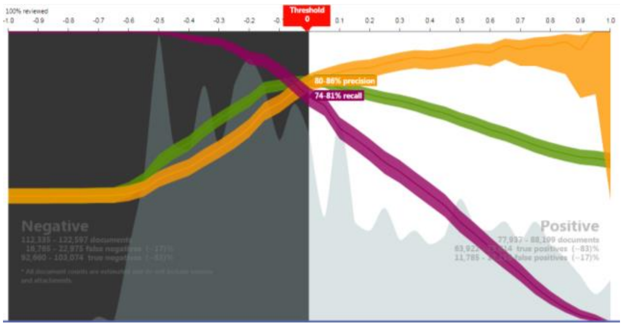
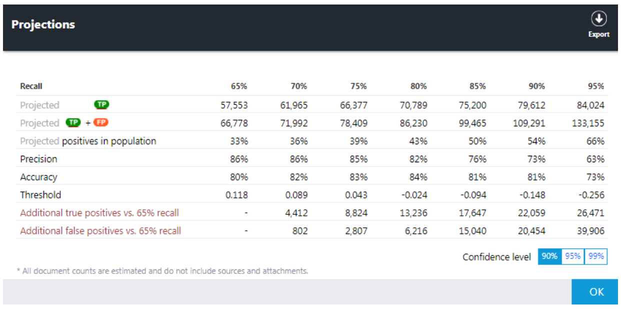


Figure 5. Ringtail predictive model.



Projections

Recall	65%	70%	75%	80%	85%	90%	95%
Projected TP	57,553	61,965	66,377	70,789	75,200	79,612	84,024
Projected TP + FP	66,778	71,992	78,409	86,230	99,465	109,291	133,155
Projected positives in population	33%	36%	39%	43%	50%	54%	66%
Precision	86%	86%	85%	82%	76%	73%	63%
Accuracy	80%	82%	83%	84%	81%	81%	73%
Threshold	0.118	0.089	0.043	-0.024	-0.094	-0.148	-0.256
Additional true positives vs. 65% recall	-	4,412	8,824	13,236	17,647	22,059	26,471
Additional false positives vs. 65% recall	-	802	2,807	6,216	15,040	20,454	39,906

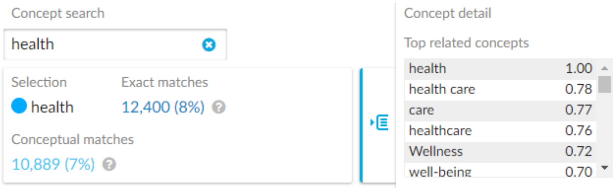
Confidence level 90% 95% 99%

* All document counts are estimated and do not include sources and attachments.

Figure 6. Ringtail predictive model projections.

2.2.4 *Luminoso*

In the case of Luminoso, a PST file must be converted to a csv format in order to be loaded into the tool. Within Luminoso, initial content analysis starts with a word cloud containing the top concepts found in the uploaded email messages. Additional concepts can be created and refined through keyword searches and related concepts (see Fig. 7). Through APIs, predictive coding capabilities can be added to the analysis process, but the team has not yet begun to work with that advanced functionality.


Concept search

health

Selection: health
Exact matches: 12,400 (8%)
Conceptual matches: 10,889 (7%)

Concept detail

Top related concepts

health	1.00
health care	0.78
care	0.77
healthcare	0.76
Wellness	0.72
well-being	0.70

Figure 7. Luminoso concepts.

2.2.5 *Open Source Tools*

Having worked with ePADD previously, the team has found a lot of built-in functionality to assist in email processing and public access. Additional exploration is planned with larger datasets from this project to compare results with the commercial tools, especially to assess the efficiency and accuracy in appraisal decisions and the identification of sensitive content in need of restriction.

The TAR Evaluation Toolkit was developed by researchers at the University of Waterloo as a simulation of a review process to estimate recall. While an initial review of this tool was not successful, more research is needed to determine if such a tool or a related tool by the same researchers, AutoTAR, might be a useful addition to existing tools used by the archival community.

3. HUMAN FACTORS

For any project involving email as archival records, there is a high probability that the records will contain personal information and other content difficult to appraise. As such, human judgement becomes a strong factor in the appraisal process. For this project, the variability inherent in human judgement may represent the single largest barrier to a sustainable workflow for the ISA. A future, sustainable process for securing and providing access to email messages having archival value will greatly depend on how standards can be defined and implemented.

3.1 Dataset Acquisition

The Director of the ISA secured support from the Office of the Governor to give the project team access to email messages previously identified as having archival value. All email selected for use by the project came from persons working in offices that report up through the Governor's office. The messages were written during a timespan from 2003 through 2010. Understandably, all of the email came from past administrations. Once approval was granted from the Office of the Governor, the IT department produced an initial set of email as PST files and transferred them to the project team on an external hard drive. The dataset was secured and then text recognition and extraction processes were applied and duplicates identified using the previously mentioned tools.

While the IT personnel were cooperative with the request, producing the dataset is not currently a priority operation for a state-level IT department. A second dataset request took significantly longer to receive and, with the recurring turnover of personnel associated with changes in government, the project team questions the sustainability of this method of securing the email. A process that involves an automated deposit of the email of senior administrators may be worth investigating for the long-term.

3.2 Tagging Content

Working with predictive coding tools means working with tools that are learning what content is most responsive to one's inquiry, often doing so using an iterative training process. The process requires human reviewers to visually analyze email messages and tag them according to pre-determined criteria indicating whether the messages are or are not responsive. Using this iterative learning process, the project team focused its initial efforts on identifying sensitive content. Once the sensitive content was identified, the tools could be used to further review and identify concepts that may be of interest to researchers. This approach has highlighted the effect that variances in human reviewers' judgment can have on the results due to the nuances of language encountered in the email messages.

The human reviewers were instructed to look for content that contained personally identifiable information or sensitive communications such as what might be found in a personnel file. Prior to any review, the team imagined other types of content that might be sensitive and should therefore be considered for tagging as restricted. Concepts such as "Family Matters" or "Endearments" were envisioned so as to flag email messages that may be fairly personal in nature. The concept "Health Concerns and Conditions" was added for communications that exposed private health information.

Through multiple iterations over days, weeks, and months the variances in sentiment among human reviewers about what should be restricted or not, as well as variances in the opinion of any one reviewer, have factored into a longer learning process for the tools. Every time a decision is made to change the criteria being used to code the documents, however subtle, the accuracy of the algorithm is affected.

The variability introduced through changes in sentiment of one tagger or through the differing notions of multiple taggers about what should be considered sensitive or restricted reduces the ability of the tools to increase reliable output of the archival review process. And yet, one can easily make the case for redefining what should be considered restricted, perhaps based loosely upon freedom of information act (FOIA) standards. In addition, many of the tools offer a means by which reviewer variances can be identified and corrected, such as using a second reviewer for each document, or through a report which highlights documents where the algorithm strongly disagrees with the reviewer's assessment. Despite the means by which these tools allow the reviewers to reassess their decisions, the challenges associated with human factors continue to be a weak link in the archival appraisal process.

4. CONCLUSION

To foster transparency and accountability in governance, researchers and the public must have access to information from government officials that provides insight into their actions and decisions. For archivists to reliably preserve large collections of digital documentation of diverse government operations, the need to leverage scalable technology is increasing. Preliminary findings from the project, *Processing Capstone Email Using Predictive Coding*, support the use of e-discovery tools to more efficiently complete archival workflows and enhance access.

4.1 Next Steps

Additional work to be done by the project includes exporting the results to preservation and public access systems, and developing an access restriction policy for electronic material. Augmentation of descriptive metadata through concept clustering is also a desired outcome. Project team members plan to ingest more email into the collections already evaluated so as to provide for a richer understanding of the operations of state government. It may be possible to use an existing predictive coding model built from the earlier ingested collection and apply it to future collections from additional people, agencies, administrations, or states, or even in an entirely different type of institution, but further evaluation is needed to make that determination.

4.2 Future Research

Evaluating tools that use predictive coding for consideration as end-user access systems and how they may be integrated into existing open source systems is a promising area for future research. Issues maintaining context with links to external resources, especially non-public intranet resources, is a concern, as is the need to re-evaluate appraisal decisions in a sustainable, justifiable, and repeatable way.

Beyond technology considerations, the successful decision-making needed to ensure the right content is reliably preserved and made appropriately accessible over time is likely to depend on at least two human factors. The first factor involves developing content-tagging protocols that can be consistently applied by human reviewers. The second factor involves determining how to develop trustworthy public-private partnerships where cultural heritage organizations may benefit from corporate investments in technology and in turn where the corporate investments can leverage broader markets. We hope to be a part of future research in these areas that will bring us all closer to having reliable and sustainable digital archives.

5. ACKNOWLEDGMENTS

The authors would like to thank the National Historical Publications and Records Commission, the Office of the Governor of the State of Illinois, the Illinois State Archives, David Joens, Amanda Myers, Thomas Habing, Dan Roth, Christopher Prom, William Maher, Jiayue Niu, Mei Mei, and Aarthi Shankar 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. 2018. Processing Capstone Email Using Predictive Coding. 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] University of Illinois Records and Information Management Services. 2016. Preserving Email Messages of Enduring Value. https://www.uillinois.edu/cio/services/rims/about_rims/projects/preserving_email_messages_of_enduring_value/.

[4] Stanford University. 2018. ePADD. https://library.stanford.edu/projects/epadd.

[5] North Carolina Department of Natural and Cultural Resources. 2018. Transforming Online Mail with Embedded Semantics (TOMES). https://www.ncdcr.gov/resources/records-management/tomes.

[6] University of Illinois Records and Information Management Services. 2016. Processing Capstone Email Using Predictive Coding. https://uofi.box.com/v/iPRES2016paper.

[7] Preserving Digital Objects With Restricted Resources (Digital POWRR). 2013. Tool Grid. http://digitalpowrr.niu.edu/digital-preservation-101/tool-grid/.

[8] Gartner, Inc. 2015. Magic Quadrant for E-Discovery Software. https://www.gartner.com/doc/3055717/magic-quadrant-ediscovery-software.

[9] Microsoft. 2018. Advanced eDiscovery. https://support.office.com/en-us/article/office-365-advanced-ediscovery-fd53438a-a760-45f6-9df4-861b50161ae4.

[10] OpenText Corp. 2018. OpenText Axcelerate. https://www.opentext.com/what-we-do/products/discovery/axcelerate.

[11] FTI Consulting, Inc. 2018. Ringtail. https://www.ringtail.com/.

[12] Luminoso. 2018. Luminoso Analytics. https://luminoso.com/products/analytics.

[13] Cormack, G. V. 2013. TAR Evaluation Toolkit. http://cormack.uwaterloo.ca/tar-toolkit/.

[14] Microsoft. 2018. Office 365 Enterprise E5. https://products.office.com/en-us/business/office-365-enterprise-e5-business-software.

[15] University of Illinois Records and Information Management Service. 2017. Tools List. https://uofi.box.com/v/CapstoneProjectToolsGrid.