The Future of Email Archives: A Report from the Task Force on Technical Approaches to Email Archives

Report Overview and Recommendations for Discussion at Digital Preservation Coalition Briefing Day

January 24, 2017

Sponsored by

THE ANDREW W. MELLON FOUNDATION

DigitalPreservationCoalition

The Future of Email Archives: A Report from the Task Force on Technical Approaches to Email Archives

---

Table of Contents from Draft Report

The Task Force on Technical Approaches to Email Archives

About the Task Force

In November 2016, The Andrew W. Mellon Foundation and the Digital Preservation Coalition announced the formation of a Task Force on Technical Approaches for Email Archives.[1] The charge of the Task Force was to construct a working agenda for the community, focusing on the following three issues: (1) articulating this technical framework, (2) suggesting how existing tools fit within this framework, and (3) beginning to identify missing elements. This report represents the outcome of our work.

For more about the Task Force, see http://www.emailarchivestaskforce.org/

Executive Committee

Christopher Prom (co-chair), University of Illinois at Urbana-Champaign

Kate Murray (co-chair), Library of Congress

Fran Baker, University of Manchester

Matthew Connelly, Columbia University

Wendy Gogel, Harvard Library

Task Force Members

Hillel Arnold, Rockefeller Archive Center

Courtney Cain, Lake Forest College

Euan Cochrane, Yale University Library

Kevin De Vorsey, National Archives and Records Administration

Glynn Edwards, Stanford Libraries

Riccardo Ferrante, Smithsonian Institution Archives

William Kilbride, Digital Preservation Coalition

Jessica Meyerson, Educopia Institute

Erin O'Meara, University of Arizona Libraries

[1] For official press release, see https://mellon.org/resources/news/articles/mellon-foundation-and-digital-preservation-coalition-sponsor-formation-task-force-email-archives/.

Michael Shallcross, University of Michigan

Joel Simpson, Artefactual Systems

Camille Tyndall Watson, North Carolina Department of Natural and Cultural Resources

Richard Whitt, Google

Julian Zbogar-Smith, Microsoft

Disclaimer

The content of this report should otherwise not be considered an official communication by the individual institutions with representatives on the Task Force.

Acknowledgements

The Task Force on Technical Approaches to Email Archives generously was supported by the Andrew Mellon Foundation's Office of Scholarly Communications. We would especially like to thank Donald J. Waters, Patricia Hswe, Kristen C. Ratanatharathorn, Tasha Garcia, Molly McGrane-Cleary, and Celia Bradley from the Mellon Foundation; the Digital Preservation Coalition William Kilbride, Sarah Middleton, and Sharon McMeekin; and project assistants Courtney Cain of Lake Forest College (formerly of University of Illinois at Urbana-Champaign) and Shreya Udhani (University of Illinois) for their contributions in managing the sources to which this report makes reference.

The Task Force gratefully acknowledges the work and contributions of Artefactual Systems; Harvard University's Grainne Reilly, Skip Kendall, and Keith Pendergrass; Preservica's Jon Tilbury; Stanford University's Josh Schneider and Peter Chan, University of Waterloo's Maura R. Grossman and Gordon V. Cormack; fwd:Everyone's Alex Krupp; the National Historical Publications and Records Commission (NHPRC)'s Nancy Melley; North Carolina Department of Natural and Cultural Resources' Kelly Eubank and Jeremy Gibson, and Sarah Koonts and the Council of State Archivists' Anne Ackerson.

6. The Path Forward: Recommendations and Next Steps

Without preserved and accessible email collections, archives and archivists are unable to fulfill their core values.[2] Responsible custody is undermined, accountability is abandoned, and, ultimately, history is imperilled. In short, the problem won't take care of itself and, the time to take action is now.

At the highest level, this report demonstrates that it is possible, but still difficult, for archival repositories to acquire, appraise, arrange, describe, preserve, and provide access to email-based collections. Repository staff must choose from a range of tools, then chain them together into often complicated workflows. While this is feasible for relatively well-sourced or tech-savvy institutions, the majority are being left behind. This is not because existing tools cannot preserve email collections, but simply because the problem is complex. The community and tools are developing but not yet fully mature. In some cases, basic research and policy decisions remain.

The challenges are clear, and some good practices have been established. Email can be preserved but the tasks ahead require active commitment and engagement from a wide range of stakeholders. Accordingly, the Task Force proposes a set of core recommendations focused in two complementary topical areas: 1) Community Development and Advocacy and 2) Tool Support, Testing, and Development. For each area, the task force lists a range of suggested activities. These include both low barrier actions, which the community can start to address immediately, as well as projects that require more planning and funding.

Community Development and Advocacy

The most important work to be done in moving email preservation forward lies simply in nurturing and fostering archives and libraries who are leading the work or with to become more fully engaged. In theory, every archive that is preserving or collecting contemporary record material has a need to collect and preserve email, making clear the need for increased knowledge, information sharing, and collaboration. These activities can, to a certain extent, be fostered through existing structures and organizations that focus on information sharing and professional development, such as the Society of American Archivists, Digital Preservation Coalition, and others. For example, a number of low-barrier activities that could be pursued

[2] Society of American Archivists, "Core Values Statement and Code of Ethics," May 2011, https://www2.archivists.org/statements/saa-core-values-statement-and-code-of-ethics.

through existing groups are described below. However, some external support and encouragement would also help the community to coalesce around certain tools and services, providing a most sustainable long term trajectory for the essential technologies necessary for email preservation. Accordingly, the Task Force also recommends a few higher-impact activities that could be pursued in partnership with organizations supporting the cultural heritage community.

Low-Barrier/Short-Term Actions

Assess Institutional Readiness for Email Collections

The community needs an assessment mechanism to help repositories evaluate institutional readiness for email acquisition, processing, preservation, and access. Understanding where functionality, staffing, and tooling are strong and where they need improvement will help institutions enhance their existing digital preservation systems and workflows.

> **Activity:** Develop a version of the NDSA Levels of Digital Preservation to address the specific needs of email and host it on a publicly accessible Web site.[3]
>
> **Planned action:** Members of the Email Task Force will put out a call for participation and convene a working group in summer 2018. Depending on institutional commitments, this work may benefit from limited external support.

Training and Skills Development

The archival community needs both training for awareness and training for competency on the core issues for archiving email. Many repositories have yet to acquire an email collection. There is a chicken and egg problem: Archivists are unlikely to solicit email until they feel competent enough to deal with the technologies and to meet concerns raised by donor or institutional partners (such as records managers or legal counsel). Put bluntly, people need to trust archives to manage email in a responsible fashion, and they to see that email preservation adds value to their organization. Some repositories have email collections in hand but need help with next steps, while others need to start preparing for its arrival.

This means that repositories should identify and train personnel who can work with large scale email collections. While some specialization is needed, the community can also train current archives and LIS staff members, leveraging existing training structures. In addition, multidisciplinary projects, such as http://www.history-lab.org/ or https://uwaterloo.ca/web-archive-group/, may offer a potential model for specialized training

[3] NDSA Levels of Digital Preservation. http://ndsa.org/activities/levels-of-digital-preservation/

and information sharing. Such groups seem more likely to succeed once basic training is in place.

> **Activity:** Develop a scalable training and workshop curriculum regarding the basics of email archives, including an overview of the issues and demonstration of available open source (and potentially proprietary) tools. A three-hour/half day session will serve as a primer to email preservation, using this report as a guide. A full day session will include tool demonstrations (perhaps recorded if needed) and active learning opportunities. Once finalized, the curriculum will be available for reuse internationally.
>
> **Planned action**: Members of the Email Task Force will develop this project for submission to iPres 2018 in Boston and will release the training materials after the meeting (assuming the session is accepted). Feedback will be incorporated into a revised version submitted for ICA 2019 in Edinburgh. Moving forward, regularized and more sustainable training could be developed for potential integration with existing curricula, such as those supported by SAA, CoSA SERI, and/or NAGARA.

Demystify Email Archiving for Collection Donors

Donors of private digital collections are often confused about the importance of including email as a documentation source as well as assurances of privacy and security.

> **Activity:** Develop a customizable template for donor agreements which details the roles and responsibilities of both the donor and institutional repository, including accessing workflows, sensitivity review, redaction capabilities, potential embargo periods, search and access.
>
> **Planned action:** Members of the Email Task Force will put out a call for participation and convene of a working group in summer 2018.

> **Activity:** Develop training videos to help archivists and donors understand ePADD's functionality in the appraisal module.
>
> **Planned Action:** Stanford University Libraries will develop and post these videos.

Maintain Assessment of Email Tools in COPTR Registry

The Email Task Force identified and analyzed common tools for email archiving. Because software tools are dynamic, with functionality added and subtracted on a regular basis, this information needs to be stored in a flexible environment with wide public access.

Activity: Move tools list to Community Owned digital Preservation Tool Registry (COPTR) public wiki.

Planned action: After publication of this report, members of the Email Task Force will contact COPTR to initiate migrating the compiled data into the registry and develop a sustainability plan to keep the information up-to-date.

Develop Format Comparison Matrix for Email Formats

If format migration is part of the preservation workflow, what are the advantages and consequences for selecting a specific target format? MBOX and EML are the de facto formats for email preservation, partially based on tool integration, but there are other options including the XML-based Email Account XML Schema (EAXS) format. A format comparison matrix will help community members understand the risks of format migration as they develop preservation planning options and institutional workflows. This could build on existing models such as the Federal Agencies Digital Guidelines Initiative (FADGI) projects which compare and contrast format options for still images and reformatted video.[4] Once completed, the matrix would be community maintained on a public resource such as NDSA or DPC.

Activity: Develop a format assessment matrix which includes information about the format's structure, standards documentation, technical metadata, header fields, expected behavior in common tools and more. This work would incorporate results from the **Test Existing Tools for Data Impact and Data Loss** project.

Higher-impact/long-term Activities

Sustain the Email Archiving Community

The momentum built as a result of the Task Force's work makes this a good time to investigate steps that can be taken to strengthen the community of institutions using email archiving tools, with an eye toward the long term. Some open-source tools relevant to archival, museum and library communities are supported on consortial models, but began their lives as research and grant-funded projects; well known examples include BitCurator, CollectionsSpace, and ArchivesSpace.[5] While the software is freely available, an institution must join the consortium if it wants to support development or receive support.

---

[4] FADGI Guidelines: File Format Comparison Projects: http://www.digitizationguidelines.gov/guidelines/File_format_compare.html

[5] BitCurator consortia is administered via Educopia and costs $2,000 annually. ArchivesSpace and CollectionsSpace are both administered by Lyrasis and they offer a sliding scale depending on the size and operating budget of your institution – annual fees range from $460 to $1,725, and $2,500 for the Leaders Circle.

The email archiving community does not seem immediately poised to adopt this model, at least not yet. Several open source email archiving tools exist, they meet different needs, and they do so differently. ePADD, for example, is a widely used tool for email archiving, particularly in collecting repositories, but its long term sustainability is not guaranteed. TOMES, which began development a bit later, is more suitable for institutional archives and is beginning to make its code available on github.[6] Code for DarcMail and EAS is not yet available online, but both projects are moving in that direction. During the Task Force discussions, members from these and other projects were keen to share information and experiences. This momentum should be encouraged, complementing the specific tool development and implementation projects discussed the following sections of this report.

> **Activity:** Representatives of the main open source development projects could collaborate on a project to define high-level functional needs for a more unified email capture, processing, and access tool. This project could also result in a proposed short term funding model, recommending support need for particular tools and services, as well as steps to build an organization dedicated to longer term support.
>
> **Actions**: Develop complete project description and seek funding.

Specification Planning for Beginning-of-lifecycle Email Tools

As noted earlier, many state governments and other large organizations use industry-developed 'email archiving' tools, or may have access to such tools as part of enterprise or cloud based systems. State archivists have noted that some changes or additions to such tools would make them much more useful in capturing, identifying and managing records for state purposes, including the capture of email from capstone accounts or manager roles, or related to particular case files.

> **Potential activity:** The community might sponsor a summit or short project, perhaps in conjunction with NASCIO, CoSA, representatives of NARA, and the academic community. Working together, the group could develop a lightweight set of functional specifications, so that email archiving tools could be used to provide better risk management, transparency, integration with other business systems, and capture of archival records bearing continuing administrative, legal, or historical value.
>
> **Planned Actions:** If there is interest in pursuing this idea, members of the task force will initiate a follow-on project and proposed statement of work, collaborating with CoSA,

[6] https://github.com/StateArchivesOfNorthCarolina

NARA, and NASCIO. Such a project may benefit from external support to convene working meetings.

Develop Criteria for Email Authenticity

More exploration and documentation is needed to test for completeness, non-alteration, and other aspects of email messages as they are moved, migrated, and processed through different points of the preservation workflow. The primary goal of such an effort is improved tooling, perhaps including, potentially, the development of a profile or schema of some variety to be used in validating the authenticity of specific properties of a message or account deemed worthwhile for judging its authenticity.

At the most basic level, the community would benefit from a common understanding of the criteria or definition of "authentic email." For example, email headers may include a variety of fields related to signature or authentication testing, which are used to indicate authenticity at point of delivery. But how much utility do such fields retain over the long term? Is email that lacks bcc or distribution list information authentic but incomplete, or is it something else? Is email more authentic when rendered in a particular piece of software within its original account context? And how can such factors be better captured to allow users to understand then interpret the layers of evidence that may provide greater certainty that a message has been unaltered?

> **Activity:** In 2012, the InsSPECT Project developed a testing process to define the significant properties of email messages, then determine whether were they were conserved when exported or migrated from a few test systems, including Thunderbird and Outlook.[7] The basic methodology was sound, but the work should be brought up to date to take cognisance of new tools and evolution in email formats (such new headers).
>
> **Planned Action:** While this work could be run somewhat in tandem with the Task Force's recommendation to **Test Existing Tools for Data Impact and Data Loss**, authenticity issues should be drawn out as a specific focus and research agenda, possibly supported by funding and through a collaboration with iSchool programs that can facilitate such work, in tandem with practitioners.

Improve the IETF RFC Standards Documentation for MBOX

[7] Gareth Knight, "Significant Properties Testing Report: Electronic Mail" (JISC, The National Archives, and Kings College London, February 12, 2010), https://web.archive.org/web/20151024134638if_/http://www.significantproperties.org.uk/email-testingreport.pdf.

Current version of the IETF RFC 4155 for MBOX does not fully describe the variations of the MBOX format. There are at least four subtypes of MBOX (MBOXO, MBOXRD, MBOXCL, MBOXCL2), which build on the common MBOX structure. Tool sets for one version are not necessary compatible with another. Clarifying the standards documentation in the RFC would help improve standardization of the format overall, enable more accurate format identification and characterization, and improve tool interoperability.

> **Activity:** Contact IETF to identify process for revising a published RFC. Contact original RFC 4155 authors and other potential contributors to form a working group for revising specification.

Improve Standards Documentation for EML

The EML format is only partially documented through IETF RFC 5322 for IMF (Internet Message Format). While IMF is defines the ASCII-text based syntax for all email messages, the EML format is a subtype of IMF used by Microsoft Outlook Express as well as other email programs such as Apple's Mail client. There is no publicly available standards documentation for the EML format although it is a common format for email archiving, including within the Harvard EAS system.

> **Activity:** Contact IETF to identify process for creating a new RFC. Contact potential contributors, including Microsoft, to form a working group for creating a public specification.

Improve Options for PDF in Email Archiving Workflows

Options to output email messages to PDF are well integrated into many common email clients. However, important header fields and other key technical metadata is often lost or concealed in the format migration. In addition, message threading and connections to attachments are terminated. Improving the technical capability of PDF software, especially software embedded in email clients, to address issues relevant to email archiving would enable simplified workflows at a large scale.

> **Activity:** Work with PDF Association, the international vendor neutral organization focused on PDF software and tools, to identify software requirements for email archiving features for the PDF format.
>
> **Planned action:** Task Force members will contact the PDF Association to start the project in Spring 2018. Based on initial conversations, the PDF Association is eager to participate.

Tool Support, Testing, and Development

Tools such as ePADD, EAS, DarcMail, and TOMES play complementary roles, meeting particular needs in collecting repositories, institutional archives, and government. There is a role for all four tools, but these projects depend largely on support from their parent institutions (and to a lesser extent, partner repositories) or funding from federal granting agencies, whose forward funding is uncertain. In addition, repositories frequently rely on commercial tools, in order to undertake specific actions to prepare or work with email. And two whole classes of industry tools that contain features of potential utility to archivists—email journaling systems and compliance/legal tools—are largely inaccessible to archivists due to their cost. Applications such as these could help immensely with two difficult tasks, capture and sensitivity review, if they were made more affordable or if open source versions were developed.

Tools need to support small and large collections alike (both collections covering many accounts and single account with many messages or attachments). In essence, email archiving tools need additional ability to scale up or down as necessary, since what is large today is not large tomorrow. The following recommendations are directed to the software development community as well as funders.

Low Barrier/Short Term Actions

Test Existing Tools for Data Impact and Data Loss

Current workflows for email archiving typically involve a common set of tools, both open source and proprietary. The impact of these tools, especially during format migration, has not been documented and evaluated. For example, are technical metadata and header fields added, lost or altered? Does the ordering of the tool chain make a difference? Does one tool perform better for a specific email format over other tools? Does one format outperform another format? How much of the envelope is retained? The first phase of this work will focus on the format migration of email messages with follow on work for renderability. The outputs of this foundational exploration will inform future work, including developing a definition for authenticity, defining a data model for email and highlighting sneed for future tool development work.

> **Activity:** Assemble varied sets of email from a range of repositories, including email that makes use of standard headers and header extensions. Move selected messages from tool to tool, comparing headers after each process. Monitor for data loss and changes, evaluating individual tools and recommending particular workflows or necessary tool improvements. The InSPECT work can be used as a model.

Planned action: In spring 2018, selected Email Task Force members will develop a project plan and apply for funding to explore the impact and effectiveness of a defined set of format migration tools on a variety of email datasets from different accounts.

Improve Format Identification, Characterization, and Validation Tools for Email Formats

The archival community needs more accurate and flexible tools to identify, characterize, and validate formats commonly used for email messages in order to increase confidence with archival workflows. It's important to integrate these capabilities within existing tools or within closely aligned applications rather than stand alone instances to promote interoperability. This would be a follow on activity to the **Test Existing Tools for Data Impact and Data Loss** project.

Activity: After the completion of the format analysis and authentication project, work to include improved options for format identification and characterization in commonly used tools including JHOVE, Siegfried and Apache TIKA.

Higher-impact/long-term

Sustaining and Integrating Existing Tools

Given the variety of ways that email can be stored and captured, as well as the need for institution and collection-specific screening needs (and the almost infinite number of ways email can be rendered or displayed to users), it is not surprising a wide range of workflows and tools are being employed. Yet, as noted in the Workflows and Implementation section, some commonalities are beginning to emerge. We see complementary approaches emerging: one focused around capturing and preserving the email of private individuals and another around institutional records, such as those of government, universities, or businesses. The differences in the tools used by these two sectors reflect both the nature of the documentation that is being preserved, as well historical trends in the cultural heritage community, where institutional archives are commonly differentiated from collecting repositories. In the short term at least, complete tool convergence is neither desirable nor necessary.

Tools such as DarcMail, ePADD, EAS, and TOMES are deserving of support that allows for their closer integration and alignment over time. This need for tool support, of course, rests in balance with community development, yet community involvement does not necessarily translate to sustainability or workflows. A concerted effort by multiple institutions, preservation software companies, funding agencies, etc. is needed to help solve the gaps in current workflows and ensure better interoperability between tools. Could they be integrated into existing consortia or tools? If not, and the trend continues of having different

organizations managing different tools, how difficult will it become for each institution to administer and justify separate expenses? Likewise, the tentative industry connections made by the Task Force suggest that perhaps some additional tool integration and development would set the stage for a set of generalized services, toward which multiple repositories might contribute.

> **Activity:** Stanford University has mentioned the need for an aggregated discovery site . This would allow uploads from multiple repositories in order to publish a greater breadth of what has been processed and would be available for research. The development of such a portal might solve one of the issues with institutions that do not have the IT bandwidth to set up a discovery server instance themselves, while also demonstrating interest in other email related hosted services**.** Stanford is investigating internal options to hosts such a site. These conversations could benefit from external support and partnerships.
>
> **Activity:** North Carolina State Archives is planning to apply for a second round of funding for TOMES, focusing on topics of core importance to the preservation of government archives and also speaking to questions that would help processing and access of personal archives.[8] TOMES is seeking to build connections to additional state and university archives, as well as collecting repositories.
>
> **Activity:** Something about EAS plans from Wendy
>
> **Activity:** Something about DarcMail plans from Ricc
>
> **Potential Actions:** Given these tool development plans, and other activities mentioned in this report, such as the need for greater interoperability and alignment, institutions that are actively developing tools would benefit from continued opportunities to collaborate. Perhaps a medium term (three to five year?) Email Archives Tool Consortium could be fostered with support of parent institutions, partner repositories, and external supporters.

Self-archiving Tool

Most of the email capture, processing, and preservation tools discussed in this report are aimed at meeting the needs of records managers, compliance officers, or archivists. Yet many people

[8] Specifically, the proposed TOMES 2.0 project would include improve flagging abilities by incorporating machine learning (to separate record from non record materials and and develop a tool that would allow archivists to prepare and deliver deliver email dissemination packets to researchers.) The project also seeks to build bridges to the ePADD project for the access packets.

have a need or desire to keep a record of their own digital footprint, in a way not unlike people of past generations, who stashed letters in a drawer, without taking the immediate step of donating it to a repository. Similarly, employees may wish to prepare a copy of their email so that successors can search and benefit from the institutional memory buried in their accounts. In other words, there is unmet demand for a service that would allow for self-directed email capture for preservation, discovery, and use, in much the same way that the Internet Archive allows people to capture websites or other resources that they find valuable. Such software could run locally or, more likely, as a web service. Data would be under the direct control of the depositor, as well as whomever he or she wishes to share it. Such a service could even be designed to allow the evental donation to a repository, via an export feature or other method.

> **Activity/Planned Actions:** If there is interest in pursuing this idea, selected Email Task Force members can develop a project plan and apply for funding to assess existing tools and to develop such a service. This may be an area for potential collaboration with industry partners or research units, or for repurposing tools such as ePADD.

Develop Standards For Tool Interoperability with a Reference Implementation

The lack of systemic interoperability between the variety of systems and tools needed for preservation action poses challenges to those engaged in email archiving. The community needs to agree on the best standards to: 1) exchange email collections through mechanisms that are secure and that maintain data integrity, 2) enrich email collections with metadata that can be operated on by other tools, and 3) maintain a comprehensive record of the chain of custody of a collection processed using multiple tools.

Many standards already exist that could be used or built upon to solve this problem. For example, PREMIS is a well adopted standard used for recording chain of custody (among other things). The Research Data Alliance has a working group developing a standard API specification for exchanging collections between repositories. Other approaches include metadata application profiles and packaging standards.

However standards only succeed when they are widely implemented and maintained. The task force therefore recommends a project to develop standards at the same time as a reference implementation of those standards built into existing tools. Developing standards with implementation partners ensures that recommendations are practical, feasible and proven to work in a real setting. A reference implementation then provides immediate value to users who can chain those tools together, while ensuring that lessons learned are incorporated into the standards and are available to the entire community.

> **Activity:** Review existing standards and identify gaps, including needs for API development. Agree on a community data model and core (or 'preferred') standards

needed for email tool interoperability. Develop enhancements to those standards (where / if necessary) to support email specific needs. Work with existing tool providers to implement those standards and demonstrate a fully functional, interoperable workflow.

Planned action: Selected Email Task Force members will develop a project plan and apply for funding to agree on core standards required to support interoperability between email archiving tools, and implement those standards in a select set of tools.

Improved Tools for Sensitivity Review

One of the most pressing needs faced faced by every repository and collection is for more powerful open source tools to automatically identify, remove, redact, and restrict personally identifiable or sensitive information—a process commonly known as sensitivity review. While there is functionality for structured classes of PII data such as Social Security numbers and phone numbers, this does not extend to less structured information such as education records (covered by FERPA[9]) or health records (covered by HIPPA[10]).

Natural language processing tools used for email review should be enhanced or complemented by machine learning software, to improve the ability of collections managers ability to identify and extract more nuanced entities from the archive. Current NLP workflows rely on named entity recognition to identify just a few data types, such as persons, corporations, places. While machine learning is not foolproof, there is evidence that use in facial recognition systems for example, that it can be highly accurate for large complex data sets, and there is no reason to think that it cannot be applied to email.[11] And the presumption that human review of messages for sensitive content is just that--a presumption. It has never been fully tested or proven, so we should not dismiss the ability of a machine to complete this task.

Allies in this need for better machine learning options are the much broader declassification and legal communities. The Public Interest Declassification Board (PIDB) published a report in 2012 asking for the President of the United States to "encourage collaboration and to determine how to employ existing technologies, and to develop and pilot new methods to modernize classification and declassification" in regards to "tagging, indexing and cross-indexing, searching, mass storage, inference, and other rules-based applications to assist declassification, access, convergence, and aggregation of media, and access by historians and

[9] Family Educational Rights and Privacy Act (FERPA): https://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html.
[10] Health Insurance Portability and Accountability Act of 1996 (HIPAA): https://www.hhs.gov/hipaa/for-professionals/security/laws-regulations/index.html.
[11] China testing facial-recognition surveillance system in Xinjiang. The Guardian. Jan 18, 2018. https://www.theguardian.com/world/2018/jan/18/china-testing-facial-recognition-surveillance-system-in-xinjiang-report.

public interest activities."[12] Similarly, recent changes in the General Data Protection Act (GDPR) in the UK, including the Right to Be Forgotten which "entitles the data subject to have the data controller erase his/her personal data, cease further dissemination of the data, and potentially have third parties halt processing of the data" will impose significant requirements for sensitivity review and redaction. [13] The archival community would do well to look at the issues raised in these groups and see them as potential sources of support for the need to build powerful open-source tools, including high-quality training sets.

The process might begin by continuing efforts to systematically test existing tools in this space, then to apply lessons in open source tools that might be made available to the archival community. How well do they work? How do they compare to studies done on TAR in the legal industry; Which tools are best, or which capabilities are mature enough for use and in what context? From here, gaps and priorities would be identified and requirements established.

> **Activity:** North Carolina State Archives is current considering work on machine learning tools (using Google's TensorFlow tools) to assist with classification and review, in a proposed TOMES 2.0 Proposal.
>
> **Activity:** The University of Illinois is assessing industry "predictive coding" tools to classify email and identify materials that should be restricted.
>
> **Proposed Action:** Selected Email Task Force members are interested in developing a project plan and apply for funding to assess existing tools and to develop requirements for an open source machine learning classification tool. This may be an area for potential collaboration with other communities interested in NLP.

[12] Transforming Classification Report. 2012. PIDB. https://www.archives.gov/declassification/pidb/recommendations/transforming-classification.html, p. 26. The PIDB board includes a founding Board Member for the Digital Public Library of America.
[13] "Key Changes with the General Data Protection Regulation," EU GDPR Portal, accessed January 20, 2018, http://eugdpr.org/key-changes.html.