Coalition for Networked Information

Strategies for Preserving Institutional and Researcher Email

Report of a CNI Executive Roundtable
Held April 11 & 12, 2018
Published September 2018

Background and Synthesis

At the Spring 2018 CNI meeting in San Diego, CA, we held two sessions of an Executive Roundtable on **Strategies for Preserving Institutional and Researcher Email.** CNI and its member institutions have been addressing this issue since the late 1990s when stewards of institutional records realized that email was replacing letters and other forms of traditional paper-based communication, and special collections and archives began to accession significant amounts of electronic mail. While institutions had policies in place for archiving institutional paper correspondence, those policies were difficult to extend to electronic mail, and both records managers and archivists struggled to meet the challenges.

At the same time, librarians and archivists are now routinely receiving very large email collections as parts of the scholarly record of individuals from their own institutions (for example prominent scholars or institutional leaders) or other individuals or organizations, such as public intellectuals, non-profit public interest organizations, or politicians; the email they are accessioning will be a key part of future special collections.

While a number of institutions have been working independently to develop policies and practices related to email archiving, a Task Force on Technical Approaches to Email Archives, sponsored by The Andrew W. Mellon Foundation and the Digital Preservation Coalition, was launched in the fall of 2016 (http://www.emailarchivestaskforce.org/). Those involved in this initiative believe that a conceptual and technical framework is needed by the library and archival community, and they also seek to capture consensus on best practices. We were fortunate to have several principals from this effort join our roundtable discussions to share their perspectives. The report from the task force was released in August 2018 and is available at https://www.clir.org/pubs/reports/pub175/.

While there are many complex technical issues involved in preserving email, including the handling of attachments and the capture of contextual information about those involved in the correspondence, it's clear the most difficult questions are policy related. How is the ubiquitous mixture of personal and more public correspondence managed (particularly given that redaction in advance seems impossible at scale)? How do we protect the privacy of those involved in email exchanges, or those incidentally mentioned? Email archives from university administrative officers are a trove of sensitive personnel, fund-raising and policy materials; faculty email typically involves students. For some special collections, there are extraordinary curatorial sensitivities

that may need to be considered in dealing with contributed email corpora. What are acceptable rates of error in redaction or filtering when dealing with email collections (which, of course, are also sensitive to who can access the collections and under what conditions)?

Clifford Lynch, CNI Executive Director, opened each roundtable by noting that he is hearing increasingly that this issue is a huge headache for institutions (that is, for both records managers and *institutional* archivists) and yet of great and evident importance in preserving the scholarly record (the broad special collections and archives perspective). In effect, there are two vectors for examining this topic. One comes out of administrative records management: legal discovery, compliance, and risk management, embedded in a set of connections between IT, records management, and the institution's general counsel and perhaps the internal and external audit apparatus. State institutions, because of the Freedom of Information Act (FOIA) and state record policies, face particular complexities and constraints in this area. Understanding this vector can be further complicated depending on whether the institution has an institutional archive unit (focused essentially on the intellectual history of the institution) that may or may not be communicating or coordinating with the unit that manages institutional legal and compliance records; obviously, the reporting lines for the archive make a considerable difference here. The hand-off between those two units can be mysterious and deeply fraught; we heard evidence that in many cases it *simply didn't exist in any meaningful way*. Unexpected catalysts like a president stepping down can suddenly frame this conflict. Surely the emails of the president should be part of the archives of the institution, but when and how they get there, how they are organized, and how they shift from governance by disclosure mechanisms used in records management (think FOIA) are not clear. Instances where the general counsel tells the university archives that full records exist, but that they may get them 50 years hence, or perhaps never, are not uncommon.

At the same time, we have a second vector: archives and library special collections that are accessioning collections of material from faculty, authors, political figures, and many others, which increasingly include a large amount of email. The most essential thing to understand is that these are being driven by *individual contributions*, not the institutional contexts of the individuals. (Indeed, as we'll discuss later, these contributions sometimes seem to almost fly in the face of the obstacles of institutional affiliation by the contributors).

Trying to figure out how to effectively accession these materials, how to appraise them, how to ingest them and how to present them to researchers, are all challenges. There are hard problems regarding managing confidentiality and redaction, multiplied by the scale and relatively uncurated nature of typical email collections. The emails of individuals typically combine the personal and the less personal: What are the practical approaches to huge collections of email? Can we do triage with automated tools? What are the acceptable levels of error and risk if we rely on automatic redaction and filtering mechanisms, and how might we mitigate these risks?

Finally, it's clear that we are approaching a technology-driven inflection point, moving from an older group of special collections donors that have (perhaps a lifetime of) email "on my laptop" (based largely on older POP technology) to a new generation that invites a special collection or archive to accession their email from the Internet, the

cloud, offering Gmail, Office 365 or other credentials to harvest this material from IMAP-based services. The simple, operational version of this is offering a set of files as opposed to offering mail login information. There's an implicit, not-yet-fully-framed conflict between email archives as the "property" of the correspondent (stored on his or her laptop) and archives held by various institutions with which the correspondent has been affiliated. This distinction is going to be enormously important.

It has become increasingly evident that email interacts with the rest of the digital environment in exceedingly complex ways. Attachments, cloud-based links, URLs, and similar structures challenge us to develop a better understanding of the scope and boundaries of email capture and archiving. Many institutions provided compelling examples of these complexities.[*]

Contextualizing email archives within global discovery systems is an important emerging issue. The Social Networks and Archival Context Cooperative (SNACC) project is of critical importance here, and various email appraisal and description tools are becoming increasingly integrated with this work. Integration efforts such as links between curation tools and platforms like ePADD or ArchivesSpace and SNACC are critical developments that must be nurtured. Being able to discover the existence of special collections or archives based on individuals or organizations documented in these collections is going to radically change the work of scholarship.

But we still don't understand how directories and catalogs of administrative roles should integrate with institutional archives and records management. Indeed, it was clear from our discussions that there's a great tension between role-based (e.g. someone serving as dean of arts and sciences during a given time period) and individual (a faculty member) perspectives on archiving. This is particularly complex because not only does correspondence respect the different roles during the period of service as an administrator, but it often transcends the chronological barriers; there's lots of correspondence about one's role as a *former* dean.

Individuals representing universities, government agencies, archives, and service providers participated in the roundtables. Some of the trends, issues, and concerns that surfaced during the conversations are noted below.

Institutional Perspectives

- Some institutions described events, such as anniversary celebrations, that prompted development of task forces or groups to think about what people will

[*] Just to be clear here; it is possible to configure standard IMAP clients to essentially store a copy of everything on the local client machine, and some people do. Indeed it would be interesting to have data about how common this is. Systems like Gmail or Office 365 offer many affordances (from spam filtering to links to documents or files stored in the same cloud system), that seem to make users less inclined to store everything locally. The questions here are more behavioral than technical, and it would be useful perhaps to have more data and less anecdote and impressions.

be able to find out about the university in the future; these discussions often include examination of the role of email in institutional history.

- Several institutions reported that records management at their institution is decentralized and, as a result, may have weak policies and perhaps even weaker implementation of what policies do exist. If individual administrative units can pick their own cloud-based mail services, this can be particularly problematic.

- It is worth noting that we are still seeing the results of a culture that was established in the 1970s, and perhaps still exists to the present day, that says that if an email message is important, it should be printed and filed, and will subsequently exist within the well-understood traditional context of paper records and correspondence. Historically, some university general counsels recommended this practice. This culture is slowly fading away.

- A number of institutions noted that they have many long-term employees who have held a variety of roles, moving from faculty to administration and back. One institution is working on a policy allowing the institution to capture snapshots of email based on the position an individual has held in the university.

- Another institution is investigating whether to harvest email from a group of administrators who have been at the university for over 20 years and sequester it in order to have it available for FOIA purposes. They intend to pilot this initiative with the president's office.

- It is difficult to develop underpinning tools in general, but even more so in institutions where a number of different email systems are in use. The ability to identify threads in an email collection is difficult. One institution is examining linked data as a potential tool to assist with identifying conversations in email collections.

- One institution described the difficulties of putting together email collections that used different platforms as a result of an institutional shift from one platform to another.

- The difficulty of migrating email from one platform or preservation environment to another is massively underestimated, as is the difficulty of aggregating email from multiple sources. Quality often suffers in these efforts.

- Even at some large research universities, the archives unit, which collects faculty materials, may not be collecting faculty email. Faculty papers, including email, are generally collected voluntarily, not as university records. One institution stated that they consider faculty email to be part of the faculty member's creative work and not part of the university record.

- One participant noted that after the death of a long-time administrator, she was charged with obtaining some of that individual's records from his computer, but the family denied access and university counsel declined to pursue.

- One institution discussed their deliberations over whether to treat faculty email differently from that of administrators and whether they could have researchers use an identifiable extension in emails so that they could easily track email for their funded projects. Another institution noted that being able to look at scientists' email could yield an understanding about important decisions in a research project.

- Email files in special collections will often need to be kept offline because of privacy issues.

- One institution described receiving a congressman's records, which included two terabytes of email messages; this collection will require a high level of security. At another institution, there has been a lot of attention to how and when email from a controversial political figure will be available. Many institutions have archives of political papers that include email.

- A number of institutions noted that collections they receive from local businesses or non-profit organizations often include email.

- Archives and special collections use a variety of tools such as Fedora, Islandora, ePADD (Stanford and the University California, San Diego), and Archivematica to ingest email. The Mellon Task Force has produced an extensive list of these. One challenge is going from a collection of tools to effective workflows at the institutional level.

- While many library staff have high-level technology skills, this is not always the case with curators in special collections (including those who negotiate deeds of gift and donations of special collections material). It was stressed that this understanding is vital in early discussions with prospective donors. Training is needed in this area.

- Some libraries will not provide any access to email collections until they have reviewed the collection, processed it, and redacted or removed sensitive materials. There is great variation in the level of risk that organizations are willing to accept, and also the level of trust (and negotiation) that they are willing to enter into with researchers who want to use the collections.

- Several institutions discussed the fact that collecting faculty email often includes student email as well (such as replies to the faculty member), which raises liability and privacy issues, including Family Educational Rights and Privacy Act (FERPA) questions. For faculty in health science settings, Health Insurance Portability and Accountability Act (HIPPA) issues may also arise.

- One institution noted the European "right to be forgotten" policy, which they believed requires institutions to redact names from records if requested. There may be emails from European correspondents in the files of North American university employees, and most institutions have not addressed this in their

policies and practices. We are not clear that these particular concerns are fully justified; many additional questions are being raised regarding the interpretation and scope of the European General Data Protection Regulation (GDPR) in the context of archival collections. It is absolutely clear that the so-called "right to be forgotten" and the GDPR are both complex and confusing legal structures that are going to continue to concern those managing archival collections. It will take a considerable period of time to develop enough case law to have much confidence in best practices. And there are other potentially relevant laws under development in various places.

- There's a great deal of interest (and hope) in machine learning algorithms and their possible application in collection filtering and redaction, perhaps in conjunction with human review, but very little data on how well this actually works.

- The move to cloud platforms like Google apps or Microsoft Office 365 is creating a new set of problems that are logical successors to the challenges of dealing with attachments; here, messages incorporate links to Google Docs, or Microsoft OneDrive, or similar storage services, with all sorts of permissions problems when they are moved to archival settings. "Attached" documents no longer accompany the e-mail they are associated with.

- There was some concern expressed about "self-curation;" that is, the ability of individuals donating email collections to edit what is donated. Cloud platforms, combined with institutional records management tools, can prevent this, if we can successfully navigate the institutional/special collections barriers.

- There are interesting questions about what to do with email attachments that are corrupted with various forms of malware, and best practices here are not well established.

- It is completely unclear what policies surround *student* email, how long this is retained, and whether the student can obtain a copy of the corpus. This has some relation to retention and access to student portfolios.

- There's a lot of interest in being able to archive email *conversations* among groups of people, as opposed to the emails retained by individuals. This is particularly relevant in trying to record multi-institutional research collaborative work, for example.

Concluding Thoughts

One enormously important point is that while we concern ourselves retrospectively with email corpora, the world moves on. Today's administrators and public figures operate in the worlds of Facebook, Twitter, Confluence, Instagram, Slack, and similar communication and collaboration platforms. We must be thinking about this today from a records management perspective, and tomorrow from an archival accessions perspective. Attempts to map challenges and concerns in records and archives/special

collections have a very strong retrospective bias of which we must be wary. This is particularly important because email invokes document and correspondence based models that are very comfortable for archivists; the newer collaboration and communications environments move very far away from these models. One speculation is that archiving may well have to be "built in" rather than an afterthought.

Often a variety of units within an institution have some role in the management of email; these include the central IT organization, records management, archives, and special collections. While currently these units often are siloed within the institution, we need to pay more attention to developing shared tools, practices, and infrastructure along with an institutional policy framework. A representative from one institution lamented that he did not know how to motivate policy development for these issues at his institution.

As one roundtable participant noted, more and more official decisions on various matters, including the university budget, are being sent via email. Policies that will ensure that we document such information for the long-term are vital.

The discussion touched on, but did not examine carefully, the complexities of records management and archival curation of attachments to email, which could include important information in documenting projects or events, but would require additional attention to formats and migration.

Representatives of the Mellon Task Force described some of the work recommended in the report, including a self-archiving tool, a tool to use in discussions with donors, and a tool for institutions to assess their readiness to preserve email.

The discussions in the groups on both days focused somewhat more on institutional records than on email collections in special collections. It will be very important for the scholarly record going forward, however, that libraries clarify their policies on email collections, including when and how processing will be done, who will have access and how, what tools will be available for searching email, and how privacy of some materials will be maintained. An effective strategy will be very important for the ability of future researchers to discover, access, and review the materials they will need for their studies.

Institutional policy is lagging in this arena; even though institutions have been using email for decades, codifying policy has been amazingly slow. One academic library stated that they began advocating for email preservation 10 years ago and moving forward required a lot of relationship building, including with the president, counsel, and IT; building trust was key.

Everything that we heard underscored the extraordinary disconnect between the institutional records management/compliance/risk management perspective and the view of archives and special collections (the two vectors we spoke of earlier). Two vignettes stand out. The first is that accession of email archives in the age of cloud-based email potentially implicates archives and special collections in circumventing security policies. For example, at the end of his or her career, a faculty member could easily move their email from a laptop to the archive of their choice. But now with cloud-based email, what does it mean if the faculty member decides that institution A will get

their email, even though they are currently at institution B? We usually think about emeritus faculty donating to the institution where they did their work, but it's not that simple. Here, would archivists at institution A use the faculty member's credentials to pull copies of mail from institution B's cloud-based mail system?

The other vignette: Modern, cloud-based email systems like Gmail or Office 365 provide superb tools for allowing institutions to ensure the integrity of the email record; a mail system administrator can essentially block deletions of email for a user (for example, in response to a lawsuit that involves discovery). Normally, this happens at the direction of general counsel. Potentially, these same tools can address concerns of archivists who fear that institutional leadership will edit and groom the email record before making it available to archives, but there's no evidence that archivists are involved in policy making with regard to the use of these tools.

Lynch noted that, in addition to institution-centric email, the roundtable conversations highlighted that we have not paid enough attention to how we document organizations and activities that are collaborative or consortial; we do not have good strategies to capture this content and this issue bears some serious consideration. It's much too easy to align along individual and institutional axes. Inter-institutional collaborations, standards collaborations, and similar activities are marginalized. The very mundane-sounding question of who should archive mailing lists, and what and how permissions should be obtained to do this, is massively important and largely overlooked. This is going to be critical to understanding large-scale scientific work and technical standards development efforts, to cite only two examples. One institution reported that they are collecting the email archives for a major project on the Internet; while the consortium that runs the project gave permission for this collecting activity, the individuals whose email is being collected have not.

It is also striking how the researcher-centric perspective is missing in these discussions. It seems reasonable that faculty members should have the ability to query or review the email corpus that reflects their career, despite the fact that they have moved several times between institutions in the course of that career. Faculty may want or need access to the email they have generated over an entire career, whether to document discoveries or provide factual background for an autobiography or other work. This is potentially an extremely valuable resource, yet it isn't available easily today to most faculty, and almost certainly requires explicit action, and perhaps some above-average level of IT sophistication, on the part of the faculty member to make it happen. This issue ties into questions about personal digital archiving practices, the shift to cloud email, tendencies to move from institutional to individual email (Gmail, for example) on the part of faculty, and institutional policies about supporting faculty as they move to other institutions. CNI will explore some of these issues in a future Executive Roundtable.