The Future of Email Archives

A Report from the Task Force on Technical Approaches for Email Archives

August 2018

Council on Library and Information Resources

The Future of Email Archives

A Report from the Task Force on Technical Approaches for Email Archives

August 2018

Sponsored by

Council on Library and Information Resources

ISBN 978-1-932326-59-8
CLIR Publication No. 175
Published by:

Council on Library and Information Resources
1707 L Street NW, Suite 650
Washington, DC 20036
Website at https://www.clir.org

Print copies are available for $20 each.
Orders may be placed through CLIR's website at https://www.clir.org/pubs/reports/pub175/

Copyright © 2018 by Council on Library and Information Resources. This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Library of Congress Cataloging in Publication Data:
Names: Task Force on Technical Approaches for Email Archives, author. | Andrew W. Mellon Foundation, sponsoring body. | Digital Preservation Coalition, sponsoring body.
Title: The future of email archives : a report from the Task Force on Technical Approaches for Email Archives, August 2018 / sponsored by The Andrew W. Mellon Foundation and Digital Preservation Coalition.
Description: Washington, DC : Council on Library and Information Resources, [2018] | Series: CLIR publication ; no. 175 | "August 2018." | Includes bibliographical references and index.
Identifiers: LCCN 2018027625 | ISBN 9781932326598 (pbk. : alk. paper)
Subjects: LCSH: Electronic mail messages. | Electronic records. | Digital preservation.
Classification: LCC CD974.4 .T37 2018 | DDC 004.692--dc23
LC record available at https://lccn.loc.gov/2018027625

Cover illustration: faithie/Shuttertock.com

Contents

Acknowledgments vi
The Task Force on Technical Approaches for Email Archives vii
Executive Summary 1
1. The Untapped Potential of Email Archives 4
1.1 Email as the Story Keeper 5
1.2 How Email Archives Are Different 7
1.3 Harnessing Technology 8
1.4 Adapting Archival Practices 9
2. The Email Stewardship Lifecycle 11
2.1 Email as Organizational Records 11
2.2 Email from Personal Records and Donated Materials 12
2.3 Email Lifecycle Stages 13
2.3.1 Creation and Use 14
2.3.2 Appraisal and Selection 16
2.3.3 Acquisition 18
2.3.4 Archival Processing 19
2.3.5 Preservation 20
2.3.6 Discovery and Access for Research 20
3. Email as a Documentary Technology 22
3.1 Defining Email 22
3.2 System Architecture 23
3.2.1 Architectural Characteristics 24
3.2.2 User Features 25
3.2.3 Operational and Administrative Features 25
3.2.4 Email Message Data Model 26
3.2.5 Message Components 28
3.3 Accounts 29
3.4 Data Transmission Model 30
3.5 Vulnerabilities of Email 31
3.6 Beyond the ASCII Message: Additional Components 31
3.6.1 Attachments 32
3.6.2 Links and Resources Outside the Message 33
3.6.3 Signature Blocks 33
4. Current Services and Trends 34
4.1 The Evolving Email Ecosystem 34
4.1.1 Abuse, Abuse Prevention, Security, and Deliverability 34
4.1.2 Marketing and eCommerce Services 36
4.1.3 Consumer Email Services 37
4.1.4 Enterprise Email Services and Operations 38
4.1.5 Email Storage, Compliance, and Records Management 39
4.1.6 Compliance and Legal Tools 41

4.2 Challenges for Repositories 42
4.2.1 Capturing Email 42
4.2.2 Ensuring Authenticity 46
4.2.3 Tracking Processing and Preservation Actions 47
4.2.4 Preserving Attachments and Linked Content 49
4.2.5 Ensuring Security and Privacy 53
4.2.6 Processing High-Volume or Numerous Collections 54
5. Potential Solutions and Sample Workflows 57
5.1 Preservation Strategies 57
5.1.1 Bit-Level Preservation 58
5.1.2 Migration 58
5.1.3 Emulation 60
5.2 Interoperability to Support Flexible Workflow Design 61
5.2.1 Processing Functionality Across Multiple Tools 62
5.2.2 Developing a Community Data Model 63
5.2.3 Defining Format Requirements 64
5.2.4 APIs and Interoperability 65
5.3 Workflows and Implementation Scenarios 68
5.3.1 Bit-Level Preservation Workflow Scenario 68
5.3.2 Migration Workflow Scenarios 69
5.3.3 Emulation Workflow Scenario 74
6. The Path Forward: Recommendations and Next Steps 76
6.1 Community Development and Advocacy 77
6.1.1 Low-Barrier/Short-Term Actions 77
6.1.2 High-Impact/Long-Term Activities 80
6.2 Tool Support, Testing, and Development 83
6.2.1 Low-Barrier/Short-Term Actions 83
6.2.2 High-Impact/Long-Term Activities 84

Appendix A: Automating System Processes 90
Appendix B: Email Tools for Libraries, Archives, and Museums 91
Archivematica 91
DArcMail (Digital Archive Mail System) 93
EAS (Electronic Archiving System) 94
ePADD (Email: Process, Appraise, Discover, Deliver) 96
Preservica Standard Edition 99
TOMES Tool (Transforming Online Mail with Embedded Semantics) 100
Appendix C: Email Preservation Research Projects 102
Archiving Email Symposium 102
Carcanet Press Email Preservation Project 102
CERP (Collaborative Electronic Records Project) 103

DAVID (Digital Archiving in Flemish Institutions and Administrations) 104
Kaine Email Project@LVA 104
MeMail (Email Preservation at the University of Michigan) 105
PeDALS (Persistent Digital Archives and Library System) 105
TOMES (Transforming Online Mail with Embedded Semantics) 106

Appendix D: Reference List 107

Table of Figures

Figure 1: Email lifecycle stages 15
Figure 2: Email message data model 27
Figure 3: Bit-level preservation basic workflow 68
Figure 4: Harvard Library migration workflow scenario 70
Figure 5: Stanford Libraries migration workflow scenario 71
Figure 6: Smithsonian Institution Archives XML migration workflow scenario 73
Figure 7: Emulation workflow example: accessing a disk image of email and attachments 75

Acknowledgments

The Task Force on Technical Approaches for Email Archives was supported generously by The Andrew W. 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; William Kilbride, Sarah Middleton, and Sharon McMeekin from the Digital Preservation Coalition; as well as Courtney Cain (Lake Forest College) and Shreya Udhani (University of Illinois at Urbana-Champaign), for their contributions in managing the sources to which this report makes reference.

The task force also gratefully acknowledges the work and contributions of Artefactual Systems, which supported the participation of Joel Simpson; 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; the North Carolina Department of Natural and Cultural Resources' Kelly Eubank, Jeremy Gibson, and Sarah Koonts; and the Council of State Archivists' Anne Ackerson.

We would like to give a special word of thanks to participants in the two briefing days organized by the Digital Preservation Coalition, as well as to Lise Jaillant, who allowed us to present our work at the "After the Digital Revolution" workshop. The attendees at these meetings provided helpful feedback, which we attempted to incorporate into the report. Any errors and omissions that remain are, of course, our responsibility.

The Task Force on Technical Approaches for Email Archives

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, and this report represents the outcome of our work. In our report, we articulate a conceptual and technical framework in which current efforts to preserve email can operate not as competing solutions, but as elements of an interoperable toolkit, and we identify missing elements and areas for additional community growth. For more information 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
Michael Shallcross, University of Michigan
Joel Simpson, Artefactual Systems
Camille Tyndall Watson, State Archives of North Carolina
Richard Whitt, Google
Julian Zbogar-Smith, Microsoft

Disclaimer

The content of this report should not be considered an official communication by the institutions with representatives on the task force.

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

Executive Summary

Email has come a long way since 1971, when Ray Tomlinson sent himself a simple test message: "something like QWERTYUIOP." By the late 1980s and early 1990s, it began supplanting paper-based personal and business correspondence in people's work and personal lives. An instantaneous means to send private messages and thoughts, email doesn't just document digital life; it documents life itself.

Email is a story keeper and a storyteller; more than 2.6 billion people currently use email, and on an average day, 215 billion messages are sent and received. Behind the daily chatter, email evidence accumulates, and the future historian bides her time, awaiting the day when she can sift through the email archives, piecing together tomorrow's histories of today.

By the specific actions that archives and libraries take today, they can capture, preserve, and provide access to the evidence that email holds. Yet to date, relatively few archival programs have taken that leap in a systematic way. Part of the problem is complexity. Email is not one thing, but a complicated interaction of technical subsystems for composition, transport, viewing, and storage. Archiving email involves multiple processes. Archivists must build trust with donors, appraise collections, capture them from many locations, process email records, meet privacy and legal considerations, preserve messages and attachments, and facilitate access.

While email archiving is still an emerging practice, this report demonstrates that archives are beginning to gain ground in approaching this most complex of problems. Some choose a simple ingest-and-store preservation approach, with no expectation of immediate usability. Others use emulation, allowing researchers to interact with email in its native environment. The most popular approach migrates and normalizes email to standards-based targets. Each of these approaches, which are not exclusive to one another and can be used in combination, has advantages and disadvantages. What they all share is intricacy. Email preservation is doable, but not yet done by enough archives to achieve our shared community goal to preserve correspondence, as we did for the paper-based archives that have facilitated untold historical insights.

If we wish to change that, interoperability is key. Just as the protocols that define the email environment are heavily standardized to facilitate interoperability across the diverse landscape of email, so too must the tools to preserve email be able to interact with one another across the lifecycle. A core set of tools, both commercial and open source, are in use within the cultural heritage community. In some cases, they need a little boost, especially to ensure more accurate and intuitive search, retrieval, and—when appropriate—removal and redaction when working with large corpora of email data. With additional investment, application programming interfaces (APIs) and other automated processes can help us link tools together, enabling more seamless workflows.

The workflow scenarios within the report represent a variety of institutional and policy perspectives and demonstrate that it is possible, but still difficult, for archival repositories to appraise, acquire, process, preserve, and provide access to email-based collections. Repository staff must choose from a range of tools, then chain them together into often institution-specific workflows. While this is feasible for relatively well-resourced 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 difficult. The community and tools are developing but not yet fully mature. In some cases, basic research has yet to be done and policy decisions remain to be made. The wider community has a vested interest in advancing the informed policies and interoperable software tools associated with preserving email archives and making them accessible.

This report has both technical and advocacy goals. First and foremost from the technical perspective, the report seeks to (a) reexamine and assess current efforts to preserve email; (b) articulate a conceptual and technical framework in which these efforts can operate not as competing solutions, but as elements of a flexible and interoperable toolkit to be applied as needed; and (c) construct a working agenda for the community to improve and refine this technical framework, to adjust existing tools to work within this framework, and to begin filling in the missing elements.

In complement, the report also serves an advocacy role as a call to arms to push the community to take action on the defined working agenda by describing why email archiving is a compelling and sound investment for modern historical records and research. Without a significant advance in technologies, such as those related to large-scale data processing as well as automated sensitivity review, and their full integration into email processing work, it seems possible, if not likely, that large sections of the historical record will remain closed indefinitely to research, whether that is related to carrying out historical scholarship, documenting rights, or ensuring accountability and effective services.

While automation through predictive coding and the like can bring costs down, money and time aren't the only taxed resources. The community at large must look at using new techniques and approaches, and these require new skills, new technologies, and changes

to funding and governance. In short, the challenge of email archiving is much bigger than simply lacking the technology or the money to buy it.

Email represents an increasingly important part of the historical record. Preserving and ensuring access to this record are therefore central to the functions and values of archives and archivists. Until we can meet the challenges of email archiving, responsible custody is undermined, accountability is abandoned, and, ultimately, the historical record is imperiled. In short, the problem won't take care of itself, and the time to act is now.

1. The Untapped Potential of Email Archives

Every year, it becomes just a bit less likely that someone will stumble across a file cabinet of important memos or a shoe box of revealing letters. Instead, correspondence is quietly filling hard drives, mobile devices, and cloud services. And the nature of what is found in these email messages stands in stark contrast to yesterday's paper communications, with their long narratives, reasoned arguments, and sometimes personal revelations. Yet email also demonstrates all manner of potential insights. The medium encourages informality. Many threads mimic conversations, with their insider references, mysterious lacunae, and sideways revelations.

Behind all this, the future historian sits unobtrusively. She is biding her time, awaiting the day when she can sift the email archives to use them as one ingredient in a story about the past. But to make that future story possible, archivists must spend their present days in the fields and in the mills, harvesting, refining, and storing the electronic grist that we all leave behind when we hit "send."

While the death of email has been reported more than once, the truth of the matter is that email is a pervasive and important tool in our daily lives, one that documents our lives and reveals what it is like to live in the first generation of the digital era (Cerbain 2016). Researchers Ducheneaut and Bellotti conceive of it as a "habitat," stating:

> It is used for a wide range of tasks such as information management and for coordination and collaboration in organizations. Email is the place in which a great deal of work is received and delegated and is a growing portal for access to online publications and information services. It has become the place where office workers spend much if not most of their workdays (the application is always on and is often the focus of attention) (2001, 30).

While other digital communications technologies such as text and instant messaging continue to rise in use, email remains one of the most adopted forms of communication. In 2016, email had about 2.6 billion users worldwide, with that number expected to continue to rise in the next five years (Radicati Group, Inc. 2016). Its ubiquity is such that it links all parts of the globe and beyond (Oberhaus 2016). Space agencies employ email to communicate with astronauts, and NASA uses it to deliver 3D design files to the International Space Station so that tools or repair parts can be created on the spot (BSG Web Group 2017). The advantages of transporting data instantly rather than waiting for a spaceship are obvious, but this incredible ubiquity, flexibility, and speed can mask the complexity of what actually happens when a user composes, receives, or deletes a message.

The question of how archivists should go about preserving and providing access to email data should matter not just to scholars.

It should matter to every person who wants to research his family history, to every student who needs to complete her school project, to every lawyer or journalist working to tell a story, to every citizen who wants to understand the actions of government. The richness inherent in email collections will remain dark until we as a community of archivists, technologists, and scholars solve the technical issues posed by its preservation and access, and until those solutions are widely available and implemented.

This report has two main objectives:

- to assess and recommend methods by which archivists can engage with new processing, preservation, and access technologies for email collections, as well as identify gaps and recommend additional development; and
- to sound the call to arms about the need to invest resources in technological solutions that improve research, archiving, and access to email collections.

"The existence of archives of letters has been an invaluable source for historical research. The use of email in recent times makes the continued existence of this resource problematic. Establishing standard procedures for archiving email is one vital aspect of preserving a record of the present for the future."

Peter K. Bol, Carswell Professor of East Asian Languages and Civilization, and Vice Provost for Advances in Learning, Harvard University (pers. comm.)

The intended audience includes the archival community, digital preservation professionals, technologists and software developers, commercial vendors, historians and scholars, institutional administrators, and funding agencies and foundations. While the report highlights specific steps on the path to success, the broader community needs to coalesce around the defined needs and recommendations to lead the charge forward.

1.1 Email as the Story Keeper

As a major communication method, email documents the personal and public stories of the day. From family gossip to friendly chatter to institutional business decisions to government actions, all are now frequently documented in email accounts across the globe. As the *New York Times* recently stated, "precisely because it's inescapable, insecure and irresistibly convenient, email provides an almost uncomfortably intimate view into the historical record. It preserves time, location and state of mind, the what-when-where-and-who of every story we might want to dig up" (Manjoo 2017). Recent years have seen email highlighted as the source of information for personal, political, business, and academic issues in the news.

One international story from the United Kingdom details the 2014 destruction by the Crown Prosecution Service (CPS) of key email messages relating to the WikiLeaks founder, Julian Assange (MacAskill and Bowcott 2017). The impact on any legal implications of the data loss is unknown because CPS did not have any idea what was destroyed, saying: "We have no way of knowing the content of email accounts once they have been deleted."

The data set known as Gupta Leaks, comprising between 100,000 and 200,000 email messages from the Gupta family in South Africa, provides another stark example of email's value. The emails revealed how the Guptas used political influence to alter political activities in their favor and secure government contracts from the Jacob Zuma

government. Selected emails were made available for just 10 days in November 2017 by the Platform to Protect Whistleblowers in Africa (PPLAAF) "to assist with the completion of the [prosecution] inquiry." The original website where the emails were released is no longer functional, but the data set is now available to journalists on request (Organized Crime and Corruption Reporting Project 2017). Information found within the email had a profound impact on South African and international business and politics, contributing to the ouster of Zuma as president as well as the shuttering of public relations firm Bell Pottinger (Alderman 2017; Onishi 2017).

In the business arena, a Florida law firm's email system was configured to drop and permanently delete spam emails without alerting the recipient. An email containing an order to assess attorney's fees of up to $1 million was marked as spam and deleted, thereby causing the firm to miss filing an appeal on time (Weiss 2017).

On the academic front, emails from University of New Mexico (UNM) athletic director Paul Krebs shined a light on his involvement with UNM paying for donors' expenses for a controversial 2015 golf fundraiser in Scotland, one of several issues under investigation by the State Auditor's Office (Grammer 2017). UNM does not have a specific records retention policy other than pointing employees to the state administrative code. Email messages deemed to be "transitory" are not permanent records (Dyer 2017b). Krebs specifically directed staff to delete incriminating email, adding later that he wasn't aware the state of New Mexico has policies regarding the preservation and destruction of public records (Dyer 2017a).

In Canada, British Columbia Liberal Party's executive director and campaign manager Laura Miller was accused of deleting emails in early 2012 concerning the Ontario Liberals' decision to cancel two gas plants at a cost of more than CAD$1 billion. During the trial, senior civil servants disclosed that they did not delete or wipe Miller's email account—in a departure from the standard procedure of closing email accounts—because they feared the account, and that of her colleague David Livingston, may have crucial records on the cancellations of the two gas plants (Ferguson 2017).

The examples of email as the story keeper, documentarian, and arbiter can go on and on, even in government. Recent examples from across the political spectrum at both the local and national level demonstrate that email has influenced actions and policy as well as the court of public opinion (Gearan and Rucker 2017; Cheney 2017; Rosica 2017). A well-known case of email as an important data source dates back almost 30 years to the Iran–Contra Affair.[2] Senior National Security Council staff Lieutenant Colonel Oliver North, and Rear Admiral John Poindexter deleted from local storage more than 5,000 email messages detailing covert government actions in Nicaragua but copies of the emails remained on backup tape (Johnston 1990). The backup copies of the email messages, which eventually found their way to the U.S. National Archives and Records Administration

[2] For a brief history of what became known as the Iran-Contra affair, see Sabato 1998.

(NARA) as the repository for official U.S. government records, provided the documentary evidence to disclose the deeds of North and Poindexter, including the false historical chronology that they authored to purposely mislead the public and obscure illegal government action.[3]

1.2 How Email Archives Are Different

Email messages seem fleeting and ephemeral compared with other artifacts, such as once-secret diaries, photographs, and personal correspondence, which can be read with little or no technological assistance. They are a sort of time capsule—dependent on technologies and system configurations as well as active good will to survive past their use date. In some cases, it is clear that certain paper records should be kept, and they are deposited in libraries and archives for long-term preservation. But the survival of email and digital correspondence involves human intervention as much as it requires technology. Like other digital data, it is produced in high volume, and it is hard to decide what merits the long-term investment of archiving. While the "information age" is producing exponentially more digital data, whether through email, social media, or cloud-based data storage, archivists and historians know only too well that large parts of our past are gone because no witness kept a record or because someone else did not want them to.

The problems of email archiving arise not just because email is all too ephemeral, regardless of the importance of the information it contains, but because of how that information is organized, or disorganized, before it is captured by archivists. The decisions made or not made by those who send, receive, and manage email ultimately determine if and how it can be made available to future researchers.

When researchers now go to work in conventional archives, they are accustomed to finding archivists who are deeply knowledgeable about the collections. Archivists strive to preserve the papers and files in the same order they were arranged by the people who produced them. They then create finding aids that describe the scope and content of each collection, often down to the individual file folder. Not only do these guides allow researchers to identify the records that are most likely to be relevant, but the way this knowledge is categorized also reveals the "official mind" of the organizations that produced it. And the files themselves, including their arrangement in boxes, frequently lead to unexpected discoveries, revealing connections that otherwise might not have occurred to the individual researcher.

Electronic records in archival repositories, especially email messages, are fundamentally different. Traditional paper-based series of correspondence are often uniform in their contents and structure, whereas email collections include both formal and informal communications, mass mailings from listservs, and even unsolicited

[3] A copy of the historical chronology can be found in the papers at the Reagan Library. See Reagan Library 2017.

advertising that, when combined with the volume of messages, makes traditional records management difficult if not impossible. The organization imposed by electronic records management (ERM) systems is designed to track and store data, keep it secure, and facilitate rapid retrieval through specified search functions. As with paper-based filing systems and technology, long-term preservation and accessibility for future researchers are not typically priorities in ERM systems. These systems are built to support an organization's internal management and auditing, not to help someone from outside the organization—perhaps many years later—assess the context or significance of any particular record.

The requirements for email search functionality differ from those for other types of archival collections, even digital collections. Traditional finding aids may not be the best path for discovery of large-scale email accounts. Keyword or full-text searching is a powerful and primary method of searching an email archive but is typically limited to on-site reading room access. Moreover, this type of search does not best serve the nature of email collections in which part of the value is the threading of the messages, and the call and response of various recipients over time. Nor does it do justice to other less defined "fuzzy" searches that don't rely on specific terms. The sheer scale and volume of email collections dictate the need for archivists to supplement traditional finding aids, organizational hierarchies, and descriptive practices with technologies that enhance the context and significance of individual records.

1.3 Harnessing Technology

The risks in this new era of digital communication are therefore great, but so too are the opportunities—and this is precisely because email is "born digital" and comes with native metadata. The computer science fields of natural language processing (NLP) and machine learning are producing an array of new techniques to process large textual corpora, extract or create metadata, and thereby reveal patterns in communication streams and social networks. Techniques that treat "text as data" are creating a new, multidisciplinary field of research that is attracting a growing number of social scientists and data scientists.

Some promising tools for improved discovery within email collections include well-developed methods of named entity recognition (NER) that can identify and quantify people, places, and organizations to help researchers determine the most frequently mentioned locations or the most prominent personalities.[4] Another technique, called topic modeling, can summarize the thematic content of a collection by identifying clusters of words that tend to co-occur in the

[4] The ePADD tool incorporates custom NER functionality to enable browsing and visualization of named person, organization, and location entities within email archives using external data sets such as Wikipedia/DBpedia, Freebase, Geonames, OCLC FAST, and LC Subject Headings/LC Name Authority File (Stanford University 2018).

same email messages, which makes it possible to arrange a collection by subject, identify the most and least commonly used terms, and also see how two or more subjects might be connected in the same email (History Lab 2018).

A strong caveat to these lines of research is the difficulty that attends the potential for unfettered access to entire email accounts. Privacy concerns make it seem untenable—even if the creator of the email account has filtered out materials prior to deposit. Current practices often limit access to components of the account data or offer a mediated approach, where the archivist culls materials for research examination. This may change in the future as email archives mature and particularly as artificial intelligence and machine learning are applied to bodies of email messages. But until that time, full "text as data" research is limited to the data that have been selected by an archivist, released through legal process, or leaked through extralegal activities.

Data scientists have already used email data sets, including the Enron email data set and the Avocado Research Email Collection, for thousands of experiments on everything from analyzing communications networks to detecting social hierarchy, to identifying authors, to auto-summarizing texts (W. W. Cohen 2015; Oard et al. 2015). Preserving email collections can create tremendous opportunities for research across many different disciplines, research that will allow us to study individuals and organizations in ways that would never before have been possible. No less important, it may induce data scientists to grapple with problems that archivists cannot solve on their own, such as reviewing email for sensitive materials (those that need restriction for legal or ethical reasons) or for automatically classifying them to enhance access.

1.4 Adapting Archival Practices

Email is shaking traditional archival processes. The ubiquitous, personal nature of email—with its voluminous, casual, and off-the-cuff character—means that any given inbox can contain thousands, tens of thousands, or even hundreds of thousands of messages. They may cover any number of topics, and some of that information may prove sensitive, embarrassing, or even of legal consequence in the future. Archiving email for the long term can be fear inducing, more so than other forms of communication. People's spontaneous reaction is often an abundance of caution that takes the form of excessive message deletion, lengthy embargo periods, or simply an unwillingness to turn over the account to be archived. The onus on the archival community is to build trust through strong policies and actions with respect to accessioning, appraisal, and preservation, supported by technologies—such as those described later in this report—that enable sensitivity review, redaction, and access.

Traditional archival appraisal practices are being adapted as new forms of processing become more widely developed and implemented (Huth 2016). New forms of data mining require a rethinking of how

future researchers will use email collections. Like paper-based materials, email records that seem mundane when examined individually might be essential and revelatory when viewed in context. Machine processing and data mining open many possibilities, when wedded to the power of the principle of provenance. For example, they may help people create a comprehensive picture of social relationships and communication patterns, essential elements of history and storytelling.

The ingest and processing of email collections for long-term storage and access may involve data "wrangling," such as the parsing and extraction of metadata from message headers, identification and characterization of attachment file formats, and deployment of processes such as named entity recognition, natural language processing, or topic modeling. This is skilled work that entails a multistep process and a series of judgment calls, all of which need to be documented for a full understanding of the content in context because different decisions can change the way the data is preserved and presented to the researcher.

As the acquisition of digital materials catches up with and surpasses that of paper-based material, it is becoming more common for an institution to acquire only born-digital archives, as evidenced by some authors of this report who have been asked to appraise and preserve email-only collections, such as listservs.[5] NARA's strategic plan for 2018–2022 specifies that digital-only transfers of records will be required after 2022 (NARA 2018). The United Kingdom's Digital Lives project suggested that in the future some collecting repositories should take a sample of personal digital archives in order to provide snapshots of everyday life (John et al. 2010). The same might apply more explicitly to email archives; even spam could have research value to social or cultural historians (Brunton 2013). Finland's Digitalia Project has taken some initial steps to facilitate personal archiving: the country's citizen archivists are being encouraged to take an active role in preserving their personal email and other digital collections, employing low-cost and easy-to-use tools (Jääskeläinen, Kosonen, and Uosukainen 2017). This approach could help address issues of privilege and class bias by ensuring that email from many individuals, including those who have traditionally been underrepresented in the archival record, have a better chance of survival.

But democratizing what comes in the door of the archives doesn't mean that all email accounts will be equally accessible. Like other forms of archival content, email collections fall under institutional policy decisions that define restrictions and embargo periods. In addition, email collections are bounded by technological constraints to a greater degree than other forms of digital data commonly found in archives. The very nature of email as a digital object evolving over time and moving through multiple systems makes it a challenge to preserve and access.

[5] For example, the American Library Association Archives, which is managed by the University of Illinois, acquired the email list of the Progressive Librarians Guild. See https://archives.library.illinois.edu/alaarchon/index.php?p=digitallibrary/digitalcontent&id=361.

2. The Email Stewardship Lifecycle

For the purpose of this project, the lifecycle model for email breaks out into two broad categories: institutional records and personal records.

2.1 Email as Organizational Records

In an institutional context, email is an organizational record that documents discussions, decisions, and actions performed in the course of business. As such, it typically falls under the recordkeeping responsibilities imposed by the legal or regulatory environment of the institution, as well as the administrative and business needs of the organization (U.S. Securities and Exchange Commission 2002). In theory, email should be controlled by a well-defined records management system and appraisal will be guided by established records disposition schedules. A formalized records management program would determine what gets transferred to the institutional archives, defining the scope and disposition of email to be retained for long-term administrative value, institutional memory, or historical value.[6] However, even many large, well-resourced organizations have struggled to implement a records management apparatus for email. In many, if not most, organizations, email rarely makes its way from users' individual accounts into a general records management system. To address this problem, the U.S. National Archives introduced Capstone, a whole account approach, which is discussed in more detail later and which is being applied in other contexts, including state and university archives.

In some instances, email messages are being captured and retained outside of dedicated email applications as components, that is, as part of case files. In these situations, email may be printed or converted to a format such as PDF, then associated with other record types, including photographs and documents. Envelope and header metadata, linked content, and attachments may be disassociated from the message or lost entirely, so care should be taken to ensure that the applications and approaches used to incorporate email in case and other filing systems include sufficient information to fulfill business and legal requirements. Several strategies can meet this need: capturing email in a native format such as MBOX or EML, recording information from the header or envelope as metadata, and maintaining linked content and attachments alongside the copy of the message that is stored in the case file.

As with paper records, companies often perceive email to be a legal liability. The preservation of personal data and even business

[6] The Society of American Archivists' Glossary defines *archival records* as materials in any format, including electronic formats, that are "preserved because of the enduring value contained in the information they contain or as evidence of the functions and responsibilities of their creator" (Pearce-Moses 2005).

records is seen purely from a risk management perspective, with a presumption toward disposal—deleting email messages, just to be on the safe side, in case the information comes back to haunt in a harmful way. Data protection policies and actions often are poorly understood and overly cautious, falling back on the policy of "what you can't see can't hurt you," which is implicitly hostile toward preservation. When records no longer serve a business purpose, records management policies usually require their deletion, resulting in the loss of historical information. In fact, the loss of the information is often extremely harmful. Records management practice must draw a distinction between temporary records and those having long-term value, each having a separate retention and disposition schedule, with recommendations for the latter taking historical value into account (NARA 1997).

2.2 Email from Personal Records and Donated Materials

For personal records and donated material, the path is even less well-defined. Email as personal records is typically created and sometimes preserved by private individuals as part of their own, or their family's, digital archive, usually outside of a corporate recordkeeping structure. There are exceptions to this, of course, including donated records coming from organizations that do not actively preserve their data. These organizations, perhaps with heavy reliance on an informal or volunteer labor force as in an activist organization, may find themselves with a cache of email and other electronic records that accumulated almost by accident in the course of doing business. These types of organizational records may also come in with personal records or as donated material but are nevertheless acquired outside of a defined records management structure.

In soliciting personal email collections, archivists might consider identifying writers, scientists, politicians, and others at an early stage in their careers and build a productive working relationship with them over time—offering ongoing tailored advice on good recordkeeping before any material is transferred to the archive, because these collections won't have the supporting structure of a records management plan to rely on. This approach is also crucial for building trust since many people do not consider email to be a significant part of their personal archive; if they do, they often have concerns about privacy and data security.

Even if an archivist is unable to work with creators at an early stage, opportunities should be sought to interview them in depth about their working practices and to gather invaluable contextual information. The Digital Lives Project coined the term "enhanced curation" to describe this area of activity, which might encompass recording interviews, taking panoramic photographs of a creator's digital and physical working environment, documenting personal libraries and artifacts, as well as collecting granular information about their hardware, software, and recordkeeping practices. In this way,

archivists and donors become partners in appraisal and selection.[7]

When a collecting institution acquires the email archive of an individual by gift, purchase, or loan, a legally binding donation, purchase, or deposit agreement should be signed by both parties. Issues covered in this agreement that are pertinent to email are discussed in more detail in the supporting documents for this report (Task Force on Technical Approaches for Email Archives 2018c).

2.3 Email Lifecycle Stages

For many years, records managers and archivists have used the concept of a records lifecycle to guide their activities.[8] This concept sets out the key stages of a record's lifespan, from creation through disposal or long-term preservation to discovery, access, and use. While variants of the lifecycle model exist, they all feature similar basic stages, described here for email:

—Creation and use—account holders

- Creation including drafting, sending, and receipt of email by account holders, as well as system-generated email
- Active use of email and attachments in day-to-day business or personal communications

—Appraisal and selection—records managers, archivists, and account holders

- Donor and gift relations
- Decisions related to the disposition of the email, including disposal/deletion or transfer to an archive
- Pre-acquisition appraisal

—Acquisition—archivists

- Assuming stewardship and legal responsibility for the email, including:
  - Capture and transfer of data
  - Transfer from active to permanent record (for governmental records and institutional archives)

—Processing—archivists

- Accessioning
- Arrangement and description
- Post-acquisition appraisal and review
- Identification of restrictions and application of embargoes, redaction, etc.
- Determination of level of preservation required and available

[7] The Digital Lives Project advocated the development of a more "archivally-oriented form of PIM that embraces the entire information life cycle, and is directed at securing authentic personal digital objects and making them readily available for use and reuse by the individual creators and owners beyond the immediate future" (John et al. 2010, x).

[8] The Curation Lifecycle Model was developed by the Digital Curation Centre in the United Kingdom and has been widely adopted as a high-level model for the curation of digital material of all kinds. Users of the model may enter at any stage of the lifecycle, depending on their current area of need (Digital Curation Centre 2018).

—Preservation—archivists and technologists

- Deposit of email messages, attachments, and metadata in a preservation repository
- Preservation actions (e.g., migration, emulation)

—Discovery and access—archivists, technologists, and researchers

- Descriptions of the email collections, such as bibliographic records, available to potential researchers
- Email messages, attachments, and metadata available for researcher use or reuse[9]

The lifecycle for email differs from that for other types of digital content because there is a greater need for archivists to be involved at an earlier stage. Instead of waiting to acquire personal archives toward the end of an individual's life or after their death, archivists and curators become involved, ideally, while a creator is still living and using their email account. Although early involvement isn't always possible, it is one factor that can help ensure that a creator's personal correspondence can be preserved for posterity. For example, if archivists obtain or document passwords, they can access an account directly or through a proxy in order to undertake predeposit appraisal and preservation planning.

Figure 1 provides a schematic overview of the lifecycle stages and processes that an email judged to be of archival value might go through.

2.3.1 Creation and Use

Creation or receipt marks the first stage in the email lifecycle. As Maureen Pennock notes, "Curation and preservation begins at source. Creators and recipients of email messages are therefore the first in a chain of important users" (Pennock 2006, 15). The control any individual sender or recipient holds over the ultimate "preservability" of their email will vary considerably from one context to another. Options for email created in a business or institutional setting will likely be highly controlled while options for personal email or that of not-for-profit and community organizations might be much more flexible.

At this initial stage of the email lifecycle, the key things to consider are choice of email client and where the client stores messages; this is particularly important when using multiple platforms to access and use email. In addition, the email client often determines the file format and relationship to associated attachments.

Beyond its core function as a communication tool, email also has some subsidiary uses, reflecting the vagaries of users' personal information management practices. For example, some people use email to manage daily tasks, to book meetings and appointments, and to keep a record of contacts for friends and colleagues.

[9] Another lifecycle model was developed by the Paradigm Project, focusing on personal digital archives. It maps digital curation activities onto the traditional archival lifecycle stages. Some of the terminology is taken from the OAIS Model (Paradigm Project 2008).

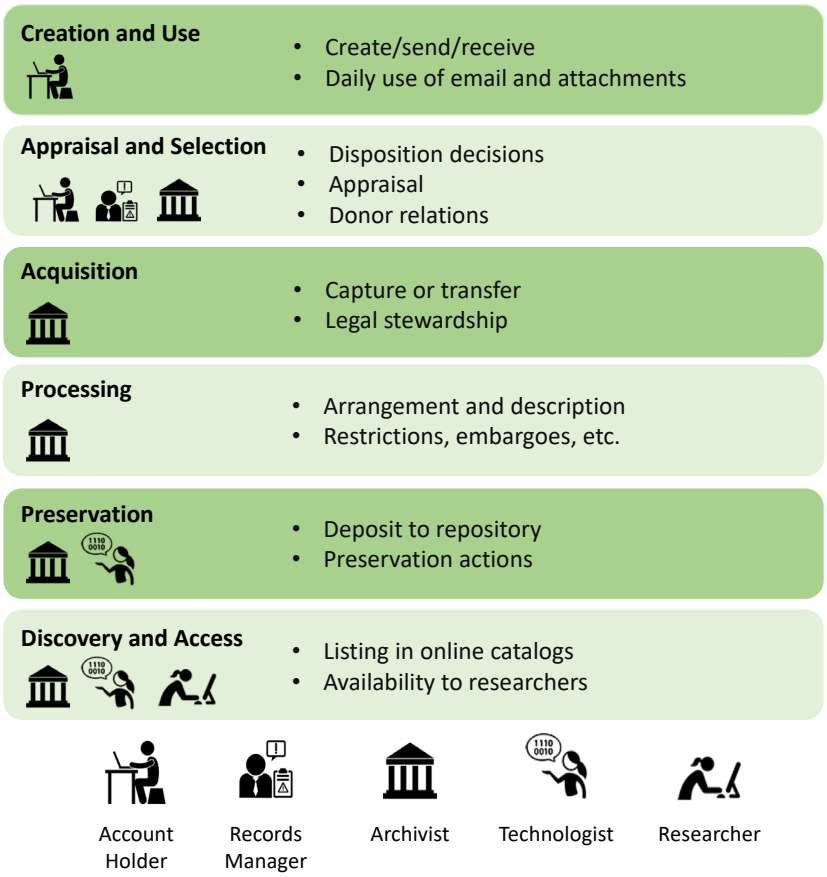

Fig. 1: Email lifecycle stages[10]

Studies have found that some people use their email account as a backup or "archiving" mechanism, emailing themselves reminders or "to do" lists, as well as important documents. They may rely on the presence of attachments if other copies of the same documents are lost; in this sense, email serves as a de facto personal archives. Some people even use their email accounts to store the ever-increasing number of passwords that they accumulate over time (John et al. 2010; Marshall 2008a, 2008b). Finally, not all email is created by humans. People's email accounts also receive a glut of auto-generated messages—such as a status change in watch folders or task management software. These messages represent the system's interaction with the email account and provide some evidence of the account owner's activities, albeit indirectly.

Authors are most engaged with the contents of their draft, inbox, and sent message folders during the time of active use. They know what information is in the messages, its context, and how it is

[10] Icons used in figure 1 are from https://thenounproject.com/. All use the Creative Commons license CC BY 3.0, except the researcher icon, which is in the public domain. Attribution for each icon is as follows:
• Account holder: Gan Khoon Lay
• Technologist: Created by Mazil from Noun Project
• Archivist: Shiva Narrthine
• Records manager: YuguDesign

organized—or not organized. Individuals' organizational habits for their inboxes vary widely, from those who create large and complex email folder structures to those who largely rely on searching or sorting for retrieval. While an individual's approach to organization is sometimes determined by the email system used, it can also reflect their personality, and archivists place some importance on the original order evident in a person's archive, believing it can add meaning to the archive's content.

2.3.2 Appraisal and Selection

Appraisal and selection is often achieved through a variety of distinct but not mutually exclusive strategies. For U.S. government agencies, an early approach to appraising email for long-term disposition was to weigh the value of each message using a records schedule or other guidelines. Some agencies may also have asked staff to identify ("tag") or move messages that they deemed worth retaining to relevant project folders. Depending on the agency, staff might be directed to print out email messages as preservation copies, to move them to another folder, or to copy them to an online system.

In 2013, NARA developed an alternative to this time-consuming and difficult-to-enforce approach with the idea of appraising entire accounts, based on an individual's role in the institution. In this alternative, called the Capstone Approach (NARA 2013a), the email accounts of identified record creators are identified for long-term or permanent retention, narrowing the focus and reducing the resources required to comply with recordkeeping guidelines. Initially suggested for U.S. federal agencies with record schedules for deposit at NARA, the Capstone Approach has been implemented in other settings as well, such as state and university archives (NARA 2013a). It specifies that the email of select individuals (such as the secretary of a federal agency and his or her direct reports, a university president or dean, or the CEO of a multinational organization) should be captured and archived. The argument for this approach is that (1) most of the email messages created and received by these individuals are official records by definition, (2) the degree of evidentiary value varies from message to message, and (3) the sheer volume of email involving individuals in these roles makes a message-by-message approach impractical, if not infeasible (U.S. Government Accountability Office 2008). When messages related to a group or project do not meet the Capstone criteria but are deemed important enough to preserve, organizations might direct staff to place those messages in a specific folder.

Sometimes a pre-acquisition appraisal is done to determine the presence of sensitive, private, or restricted information using tools such as ePADD, BitCurator, and FTK (Forensic Toolkit), which allow archivists, and even donors, to review and flag materials to be redacted or embargoed before accessioning. While searches for Social Security numbers, phone numbers, and other structured data or names are fairly straightforward, fuzzy searches for sensitive topics or knowledge gleaned from combining different data results

are much more challenging and complex endeavors. They require sophisticated or time-consuming approaches. Institutions may opt to address concerns after transfer to the archives, when time constraints are less pressing, but the need for tools to assist with sensitive message identification is great.

Appraisal is time- and labor-intensive, so archives often appraise at the collection level only and leave detailed appraisal work until after acquisition. In some cases, an institution may decide that the resources required to undertake any appraisal at all are unjustified, when balanced against the low cost of large-capacity digital storage (The National Archives 2016). It has been argued that in a historical or cultural context, where email hasn't been subject to legal recordkeeping requirements, "there is increased merit in the 'keep everything' approach" (Pennock 2006, 18). In the case of personal archives—like those of writers—appraisal may, in any case, be minimal. Sometimes appraisal is limited by technology approaches. If the institution is taking a forensic disk image of a creator's computer—which enables a digital archivist or curator to capture the individual's entire desktop environment, including but not limited to client application email accounts (but not those in web-based interfaces)—appraisal is possible only after acquisition.

As with individuals curating their own email, it may be feasible for an archivist to undertake a basic level of pre-acquisition appraisal simply using the native functionality of the creator's email application. Another approach that has been suggested is to retain the sent items folder only, as this will typically capture all items that were important enough to be replied to or forwarded and will retain both sides of a thread.[11] Similarly, the kind of functional appraisal that is applied to larger organizations can work for individuals or small institutions, e.g., retaining records based on their documentation of particular functions, projects, or areas of work.

As part of appraisal, there are many approaches to the disposition of records (that is, their final destruction designation for transfer to an archival institution). Institutional archives, which document a parent organization, may implement records disposition schedules to define what content should be retained in the creating offices, and for how long, before destruction or transfer. Email messages or accounts not deemed to be official or historical records might be disposed of immediately or be temporarily retained for business continuity purposes. An institution's email management procedures affect how this occurs and may impact the thoroughness of its recordkeeping program. Recordkeeping obligations to other organizations, such as external funders, should be included in records disposition schedule development.

Ideally, an organization's members and IT staff will collaborate to manage email messages in accord with well-considered record management schedules and guidance. Of course, less formal

[11] This is the strategy used by the New York Philharmonic archives. However, one study of an art museum's archive found it ineffective in capturing all the messages classified as "significant" (Cocciolo 2016).

processes sometimes predominate. For example, IT departments may periodically delete very old email messages in staff accounts or may temporarily block email traffic to large accounts until the owner deletes enough messages to reduce the account size to meet their organizational standard. Both approaches create opportunities for official or historically valuable communication to be lost. On the other hand, an organization may keep all of its email, including that of former staff, because it lacks a disposition schedule or the means to carry out the disposition stipulated in its schedule.

In the case of personal digital archives that are being acquired by a collecting institution, the "disposition" stage of the lifecycle refers to the transfer of material selected for acquisition following initial surveying and appraisal. Usually the material transferred will have been identified as holding historical value and earmarked for long-term preservation. Depending on the extent of the email concerned, the preserved email might include individual folders or entire email accounts. At this point, sensitive information identified by forensic tools or materials without long-term research value may be securely deleted.

For individuals curating their own digital archives, this point of the stewardship lifecycle represents the stage at which decisions are made about whether to preserve some or all of one's email for the future and how to dispose of the rest. At this stage, archivists must secure ownership of the email corpus under a deed of gift or other appropriate legal instrument. Ideally, the deed will note access restrictions or disposal requirements that the donor has negotiated with the archives.

2.3.3 Acquisition

Acquisition is the point in the lifecycle at which email identified for long-term preservation moves to the custody of the archives.

At an institution with an electronic records management system in place, email messages classified as records might be captured in the system at the point of creation; otherwise, they may be exported and stored on shared network drives according to a defined records schedule. Transfer to the archive is likely to be made using physical media such as removable hard drives or digitally by secure FTP.

Where no formal records management system or policy exists, the process of transfer may take place when an employee changes roles or leaves the organization, in which case the process is more akin to the acquisition of an individual's email by a collecting institution. At this time, email may be exported from the email client for transfer, or the institution may choose to make a forensic disk image of the creator's entire PC, capturing the email in the context of the desktop environment.

Whether transferring official email from live accounts to the institutional archive or acquiring the archive of an individual from outside of the institution, security and authenticity during the transfer process are primary concerns. Email may contain highly sensitive personal information, so the transfer mechanism needs to be totally secure; any removable media should always be encrypted. On

arrival in the archive, email may initially be stored on a quarantined computer, isolated from the network, until a sensitivity review and more detailed appraisal have been carried out; some institutions also process and preserve email on a separate, secure network, even after post-acquisition appraisal has taken place. Whatever the means of transfer, the email should be subjected to an integrity and health check on arrival in the archives. This includes running a fixity check and virus check on both the email messages and attachments.

2.3.4 Archival Processing

Archival processing takes place after email has been transferred to the archives, either from a department within the organization or, in the case of a donor gift or purchase, from an external party.

An accession record will normally be created at this stage to capture key information about the email and whether it forms an accession in its own right or is part of a larger accession of digital or hybrid archive material. This is also the point in the lifecycle when a more detailed appraisal is generally undertaken, refining any pre-acquisition appraisal that has already taken place. Together with a sensitivity review, such an appraisal is likely to combine technology-assisted review with manual input by the curator, based on scanning sampled messages (for an example, see The National Archives 2016).

When a detailed review is impractical, some institutions retain entire email accounts on a restricted basis—that is, they remain closed until requested by a researcher, at which point the relevant set of email messages will be checked for sensitive data before release. In other cases, a record series (including email and non-digital formats) may be embargoed for a set period of time as a matter of policy. Since such content is not accessible to the public in any form, there will be little immediate need for detailed review. The Smithsonian Institution Archives, for example, collects the email of key organization record creators such as the secretary, the under secretaries, and museum directors. These records are embargoed for 15 years from the point of acquisition. Other institutions have even longer embargo periods (Ferrante 2015). At Harvard University, access to faculty archives (formerly called "faculty papers") is governed by individual deeds of gift; university policy restricts access to university administrative records (analog and digital) for a period of 50 years from the date of their creation; and university records pertaining to individuals, including student and employee records, are closed for a minimum of 80 years (Harvard University Archives 2018). Princeton University Archives collects email of university administrators, and as an institutional record, it is closed for 40 years per their access policy (Princeton University Department of Rare Books and Special Collections 2018).

Other activities at this stage of processing include using tools to validate and characterize a body of archival email and attachments—extracting metadata about file formats and versions; dates of creation, transmission, or modification; authors; file size; and more to ensure the authenticity and integrity of an email archive.

2.3.5 Preservation

The preservation stage in the lifecycle represents the deposit (or "ingest") of email messages, attachments, and metadata into a preservation repository. Ingest and preservation raise several email-specific questions:

- Should the archives treat the whole group of related email messages as a single information package, or consider each message as a distinct information package?
- Should email attachments be maintained as originally received, or should they be separated, with an explicitly maintained relationship between each message and its attachments?
- How will the repository document relationships between email messages, message threads, and other groupings, such as folders?
- How will the archives implement its preservation strategy, to take cognizance of email?

Regardless of whether the decision is to treat each message as a Submission Information Package (SIP) or to treat a whole account as a single SIP, the resulting Archival Information Package (AIP) should include the email in its original format and in a normalized or preservation format. The metadata, subject and sender logs, associated attachments, if separated, and fixity data would be maintained in a database used to manage the AIPs and ideally stored with the AIP emails as supplemental files. AIPs would then be placed into a trustworthy digital repository.

At the point that a preservation intervention is required to sustain access, the preservation actions, agents, and additional data, such as those specified in the PREMIS standard, will be added to the metadata already gathered to ensure a complete set of provenance and authenticity documentation and to inform future preservation actions. The resulting preserved email will be part of a new AIP, along with the original format acquired (email and attachments) and the expanded set of metadata. The new AIP can replace the original AIP in the trusted digital repository.

2.3.6 Discovery and Access for Research

Archival email provides a key resource for the researchers, students, and family and community historians of the future. It has potential for use in many disciplines and fields—not only in traditional research areas, but also in emergent and innovative avenues of investigation.

At the most basic level, those interested in the personal archives of individuals will want to access their email correspondence as well as their letters. In addition, users will usually wish to see how their email fits into the context of their archive as a whole, whether that is entirely digital or both hard copy and digital components. For this type of research, the traditional archival finding aid is still likely to provide a useful way into an archive. Researchers interviewed as part of the Carcanet Press Email Preservation Project, for instance, valued the finding aid as a way of intellectually uniting the analog

and digital components of a hybrid archive (Baker 2015, 221). A body of email with its associated digital metadata offers the opportunity to automatically populate certain elements of a finding aid: email header information can be extracted to provide covering dates and to list senders and recipients, while text mining and semantic analysis can highlight concepts or names, which might then be pulled into the finding aid as access points. The user may then be offered the choice to move from the finding aid to the archival email itself, which might be stored in digital repository software and delivered through a separate interface where users can carry out more granular searching, filtering, and browsing.

Some researchers are primarily interested in the content of the email in a person's or organization's archive and may have no strong views about how it is presented to them—although they may wish to encounter it in a form that at least approximates the way email messages are displayed in a client or web application (Baker 2014).[12]

Others may wish for a more immersive experience and be keen to explore archival email as it would have appeared on the original creator's desktop, as with the Salman Rushdie Archive at Emory University.[13] Using emulation to recreate how an email account would originally have looked and behaved can facilitate this kind of access. It may also require researchers of the future to master the use of old hardware and software in the same way that today's historians learn about diplomatics and paleography to understand and interpret early correspondence.

The ability to extract key metadata from email also provides scope for replacing or augmenting the traditional finding aid with multiple access points into an email archive, including visualizations based on metadata or email header information.

Issues of copyright, data protection, privacy, and sensitivity may prevent the wholesale release of email for use and reuse. In the case of email held by archival institutions and libraries, it is necessary for curators and archivists both to secure copyright permission and to check email for sensitive data before it is made available remotely. This is difficult to achieve at scale, especially when thousands of third-party rights holders can be represented in any single email account. Tools such as ePADD's Discovery module provide scope for making redacted versions of emails available remotely (displaying only "entity" metadata such as correspondents, locations, and organizations), but at present many institutions are making only the full text of archival emails available on-site in a mediated environment (Owens 2014).

---

[12] Researchers consulted as part of the Carcanet Press Email Preservation Project expressed a preference for an interface that presented email in a familiar way—with to and cc fields, date, subject line, and so on—but in a way that is deliberately neutral rather than attempting to artificially reproduce something that looked "authentic" (Baker 2014, 28). Some repository software—such as Preservica—also presents email in this way.

[13] For an overview of the researcher perspective of the emulated experience, see Rockmore 2014.

Every institution that acquires email as part of its archival holdings must give thought to how its end users might wish to access and use such archives both now and in the future. The institution must also consider whether the originally acquired form should be retained in addition to the fully processed form, considering the potential implications this raises for future users who may wish to judge the authenticity and completeness of the record.

3. Email as a Documentary Technology

The primary objective of this report is to assess and recommend methods by which archivists can engage with new processing, preservation, and access technologies, and to identify gap areas that can be remediated through the development of additional tools or frameworks. Meeting these objectives requires clearly defining and understanding exactly what email is and how it works.

3.1 Defining Email

Email is both one thing and many things. It's an individual message; it's a collective noun for all the messages in a mailbox; it's an active verb ("I'll email you later today"). But behind each of these uses, email is the system that creates, distributes, and receives messages according to the rules of a defined, extensible set of standards.

Email originated as a relatively simple method of exchanging messages composed of structured text between networked computers, but it has evolved to support a wide range of functionality. It can now operate over many different networks and protocols. Of course, most people use email for the simple task of sending and receiving text-based messages and attachments. But personal information management (PIM) programs often use email to manage calendars and track contacts; it is also common for telephone systems to provide access to voicemail as an emailed audio file. Task and project management systems closely integrate with email, sending and receiving comments, responses, and to-do items. Authoring systems (for example, the Google Doc on which this report was collaboratively written) use email to send and receive comments. Social media services such as Twitter and Facebook can leave extensive traces in email, and content management systems are sometimes extended to email database backups to an account on a daily or weekly basis. Systems may integrate so closely with email that it is difficult to define where one tool ends and another begins.

At its heart, email is a transaction whereby a sender transmits a message to a recipient. For the process to complete, the recipient must be able to understand the message and, if appropriate, respond to the sender in turn. In the case of email, we can take the "message" component to mean the actual email message, but it might also include attachments, as well as embedded links to external content and other features.

The website of this task force provides a guide to what makes email so universal: standards (Task Force on Technical Approaches for Email Archives 2018b). Yet reading these documents can be frustrating if the goal is to understand how email works. Terms are often used in an inconsistent fashion. Both *address* and *mailbox*, for example, refer to the destination to which a message is sent. But as stated in RFC 5321 2.3.1: "The two terms are typically used interchangeably unless the distinction between the location in which mail is placed (the mailbox) and a reference to it (the address) is important" (Klensin 2008, 15).

In archival terms, a complete record is generally agreed to be one that includes sufficient content, context, and structure to ensure that its information can be accessed, understood, and preserved for as long as necessary, and that its value as evidence has been maintained. If the goal is to preserve email as a record of a particular action, the person or organization responsible for preserving it must understand the components of a message and account, how they relate to one another, how systems store them, and what is retained (or not retained) depending on the method used to capture messages. Institutions should ensure that their systems, training, and policies are developed with an understanding of what a complete record means to them. Based on this understanding, they should develop policies and processes that help them capture sufficient information to preserve and provide access to email records.

Several factors make this a challenging task. For example, the standards do not require that the data used to actually transmit a message (the "envelope") be recorded exactly or completely in the metadata of the message itself (the "header"). Some applications support non-standard metadata, such as tags or substitute nicknames, obscuring addresses or people's proper names. Additional data can be lost or decontextualized on export. And many messages lack information defining an account owner's job title or role. Records management and preservation activities can also affect the appearance and behavior of email.

3.2 System Architecture

Many applications are used to compose, read, organize, and maintain collections of email, and these applications take different approaches to email storage. Many of email's intended features give rise to its technical complexity. Accordingly, a sound understanding of its architecture helps us appreciate and interpret the specific preservation challenges that it occasions.

Prior work, particularly in the technical literature, has typically described what email systems do, with a parallel explanation of how the entire ecosystem works. The InterPares *Keeping and Preserving Email* report, for instance, provides a good overview of email functionality (Pontevolpe and Salsa 2009). Crocker, similarly, provides an excellent overview of the email system, combining a functional explanation with a description of system design (Crocker 2008). This section of

the report aims to provide a more direct and concise summary of what email does, focusing particularly on the architectural features of the email ecosystem, including how the system itself has been maintained and adapted. Some attention is also given to human user features, before a description of system actors, features, and standards that are used to meet particular architectural goals and user needs.

3.2.1 Architectural Characteristics

From a technical perspective, email operates as a series of commands and responses to facilitate a series of mundane processes, not unlike a computer programming language. It is easy to imagine that clicking on a Send or Forward button is like shooting a message through a pneumatic tube directly to its intended recipient. It is more like handing it to your mail carrier, who then takes it to a post office for sorting and routing (through a set of other post offices), before it is delivered to the intended recipient. RFC 5321, the Simple Mail Transfer Protocol (SMTP), specifies the complete set of commands, including HELO (which identifies a client to a server), RCPT TO (which specifies the recipient's address), and DATA (which initiates the transfer of the actual message and any attachments) (Klensin 2008, 15; Larramo 2018). Email client software serves as an intermediary, silently performing these functions, so that users are not required to memorize a long list of cryptically named commands.

Fundamentally, the email system is designed to allow messages to be sent between different system actors, which are unknown to each other until the time of transmission. Its architectural features include the following:

- **Global addressing.** The email system is able to provide addresses that are unique across the entire system.
- **Interoperability.** No prior arrangement between participants is necessary. Standard communication protocols ensure that parties can reliably exchange messages, and format standards ensure that messages can be widely and reliably interpreted.
- **Asynchronicity.** Messages can be relayed without the need for sender and recipients to be online at the same time.
- **Redundancy.** The system continues to operate reliably even if particular components within it fail. The system is based on a principle of "best effort" delivery, avoiding the onerous complexity or cost that 100 percent reliability would demand.
- **Dispersion.** The administration and operation of the email system is distributed among many different stakeholders. While there is governance through bodies such as the Internet Engineering Task Force (IETF) and Internet Assigned Numbers Authority (IANA), the system functions without any central operational control and is a self-organizing system. For example, messages do not specify a particular delivery route.
- **Backward compatibility.** The standards process allows for regular enhancement or modification to the system, but demands that components remain compatible with older versions of standards. This allows change to be introduced incrementally and prevents errors or unexpected behavior between different components.

- **Extensibility.** The communication protocols and data standards both allow for local extensions or customizations, without requiring all participants in the system to understand (or even be aware of) those extensions. This allows participants to meet local requirements, innovate, and experiment without compromising the need for global interoperability. Extensions can be completely local or registered through the IANA.

These core features have contributed to email's widespread adoption, resilience, and (in internet terms) longevity. But on the flip side, they have also imposed trade-offs and vulnerabilities, such as complexity, spoofing, spam, and phishing.

3.2.2 User Features

Email users are the human and machine actors who send and receive email messages. While most historical interest in email would seem to coalesce around its human users, computer-generated accounts and messages also interact in the email records and should be considered as agents in the email lifecycle. But whether the agent is a person or a machine, email provides a consistent set of user features, such as creating message content; adding attachments; addressing, sending, receiving, viewing, storing, deleting, and managing messages; creating message metadata; and defining automation. The website of this task force provides more information about each of these features (Task Force on Technical Approaches for Email Archives 2018d).

3.2.3 Operational and Administrative Features

Email standards also mandate system operational and administrative features. Since these have been embedded into the design of most email servers and clients, they affect the ways in which users and archivists are able to capture and preserve messages for their long-term value. Core functionality includes the following:

- **Mailing lists.** The automated operation of mailing lists is a core feature defined by email standards. A user can send a message to a mailing list address, which will then be sent by a "mediator," which receives, aggregates, reformulates, and redistributes email messages to each list member (Crocker 2008). Header fields (i.e., metadata) support list functionality. Mailing lists often incorporate an archive feature, and all messages are often kept and organized using this metadata, but mailing list archives should not be confused with long-term preservation in an archival repository, where preservation is a core operational goal, rather than an afterthought.
- **Delivery features.** Many of the systems interacting to send, receive, and store messages use standardized functions to monitor transmission and ensure message delivery. These features underpin the "store and forward" method of routing email messages, reliably delivering email when components of the wider network or infrastructure are unavailable. System actors can send delivery status notifications (DSNs) to inform senders of delays, errors, or

In October 2016, the topic of email authentication broke into a minor news story when some senders denied having sent emails found in a cache released by WikiLeaks. Tech blogger Robert Graham used digital signatures encoded in the header to verify the authenticity of a sample of disputed messages (Graham 2016). But as Graham notes, signature verification can be completed only if the referenced certificate remains available on the originating server, which seems unlikely to be the case for any extended period of time.

successful delivery. Message disposition notifications (MDNs) indicate whether a message was read or deleted. Both DSNs and MDNs are optional but widely implemented.

- **Local administration and policy enforcement.** Email providers can support a wide range of local or custom functionality, which is implemented in accord with local policies or rules. Messages may be transformed into preferred storage or encoding formats. Viruses and spam may be segregated or deleted. Content may be monitored and profanity or objectionable content redacted. Systems may impose mailbox size limits, which in turn drive delivery features, such as sending a notification to a sender when an account is over its limit and the message is not delivered.

Operational and administrative features can and often do modify both the content of messages and the metadata (header fields). As a result, entirely new messages may be created automatically, but with data and metadata that make them appear to result from human actions. While there is great value in this functionality (and limited means of removing it, given the principle of backward compatibility in the larger system), there are clear disadvantages. Many of the preservation challenges and complexities in establishing authenticity and provenance, and capturing contextual information accurately, are a direct result of these features.

In practice, the body content of each message, as well as the visible header information, are likely to be very similar, if not identical, across systems and versions of a single message. Users have a variety of means to verify message authenticity, provided the header is intact and the referenced signature systems are still functioning. It seems unlikely that future users will be able to use these techniques in the case of materials contained in email archives, since the purpose of the signatures is to ensure delivery, not to serve as a permanent audit trail. Authentication depends on the future existence of the signature system or certificate.

3.2.4 Email Message Data Model

The standard architecture of email systems and infrastructure is possible only because these systems rely on standard data models for the message format, account address, and transactional process. The IETF publishes numerous Request for Comments (RFC) memoranda that define how email messages must be structured, how they must be transmitted, and how developers can extend the platform to support new technologies. These documents hold different statuses, reflecting their level of maturity (Bradner 1996). Many remain in draft status for years, but taken as a whole, they serve as a de facto set of controlling standards.

The main standard governing the message structure is currently RFC 5322 (Resnick 2008). This standard specifies "a syntax for text messages that are sent between computer users, within the framework of 'electronic mail' messages." The format used to store email "at rest" is not covered by this standard and is determined by

the client or server applications used, as well as user-selected settings, such as whether messages will remain on the server or will be copied to the user's device or stored in both places. Email is commonly stored in a proprietary format, then reconstructed into a format such as MBOX, PST, or EML as necessary for backup or export (hMailServer 2018; Microsoft 2005; Novell Documentation 2018; Vogel and Cazabon n.d.; Zdziarski 2008).

The email message data model (figure 2) seeks to describe the components of individual email messages and strings of messages. An email message is constructed in a manner analogous to the structure of its paper equivalents. Like a paper letter, it includes an envelope that wraps message content and attachments and includes address information to ensure delivery to the correct destination account. Because the envelope's address information includes the destination and return addresses, the fields in the message header are for reference purposes only; they are analogous to the heading on a business letter. The metadata fields in the header *probably* match what was on the envelope, but the address on the envelope is what the mail carrier/email systems actually looked at when delivering the letter/email message. Most of the envelope information is hidden from a user's view. The systems involved in its creation, transport, receipt, and storage determine how much of it is retained upon successful delivery of the message. The RFC does not mandate that any of the envelope fields be retained, but header analysis can give a certain degree of confidence in authenticity.

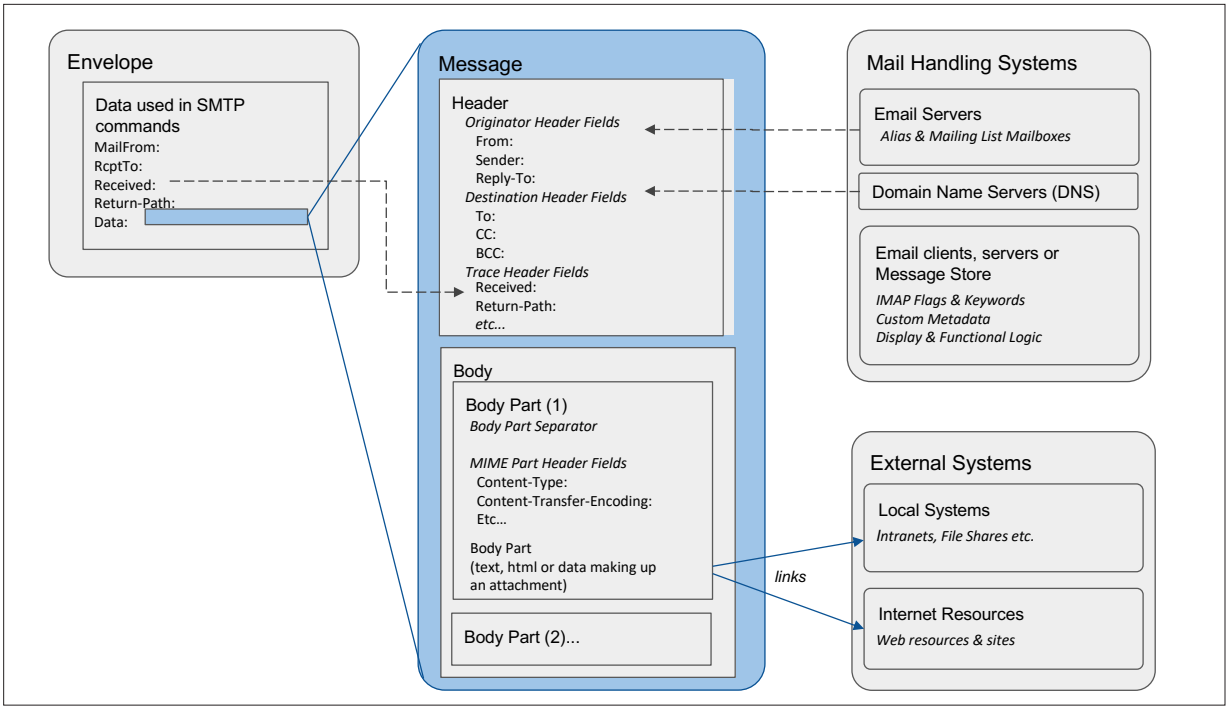

Fig. 2: Email message data model

3.2.5 Message Components

Each message that is sent or received must include certain required elements and may include certain optional ones. RFC 5321 provides specific requirements for each of the following parts of a message.

- **Envelope:** defined in RFC 5321, section 2.3.1 (Klensin 2008). An envelope consists of an originator address, one or more recipient addresses, and optional protocol extension material. Email systems typically record the information used in the envelope as metadata in the message itself (see "Header" below). However, the standards do not strictly require this.
- **Header:** defined in RFC 5321. Every email contains a collection of header fields, each consisting of a header name, a colon, and data, structured as described in the message format specification RFC 5322 (Resnick 2008). InterPARES 3 treats the header fully, with sections on identity, delivery, thread, and Multipurpose Internet Mail Extensions (MIME). Each message should be assigned an application-generated message ID as defined in RFC 2392 (Levinson 1998). Message IDs are theoretically unique and are used to relate messages to threads when referred to by the Reply-To and References header fields. The method of constructing message IDs is not clearly defined, and applications take differing approaches. They make it possible to associate messages with the application that generated them and can be used to spot forged email if the forger fails to insert a message ID consistent with the application used to generate the message. In addition, each system may or may not preserve routing and delivery information that accrues as the message is transmitted from one server to another on the way to its final destination. Somewhat akin to postmarks or passport stamps, this information is often written into stored versions of email messages with Received headers.
- **Message body:** defined in RFC 2045 (Freed and Borenstein 1996). The message body contains the message's text, HTML, attachments, and inline content (such as images).
- **Attachments (MIME):** Email attachments are addressed in the MIME document series RFC 2045, RFC 2046, and RFC 2049, which describe mechanisms for the transmission of images, audio, or other sorts of structured data in email.
- **Data model extensions:** While the core concepts of envelope, header, header fields, body, and body parts are closely defined, RFC 5321 sets out an extension model so that producers of email systems have a framework for adding metadata or functionality that will not adversely impact systems that do not support those extensions (Klensin 2008). Accordingly, certain email systems support standards-based extensions, such as header fields to support digital signatures, while other systems include proprietary extensions, such as metadata added to support voicemail.

Considering an archivist's desire to preserve email as a record, the distributed nature of the email data model and the extensions gives rise to several challenges. For example, custom metadata can

be, and often is, stored by local user agents or message stores, not in the message itself. This metadata may or may not be preserved when email is captured outside the system. While some systems store email as complete messages in a representation format such as EML or MBOX, many systems store the component parts of messages (e.g., sender, recipient, date, subject) in different tables of proprietary databases and reconstruct messages into a unitary form only as needed. The extensibility of the data model also means that messages may contain metadata that ranges from widely used (and therefore well understood and documented) to rarely used (with potentially no documentation). Similarly, there is a loose association between external systems and header fields. An email address in one of the header fields may actually refer to an alias or mailing list, so the destination mailbox(es) are not listed in the message. Also, display properties and functional logic are often stored locally. Local properties may rely upon open standards (e.g., SIEVE email filter rules) or may be custom and proprietary (which is probably more common), but in both cases can affect how email is aggregated into a user interface or represented on a message-by-message basis. For example, Gmail auto-sorts "promotional" emails into a separate tab, and Apple's Mail application will identify "contacts" or "calendar events" from information within a message. Finally, body parts can contain links to external resources (e.g., web pages or images) that seem to be embedded, but are actually just displayed by reference.

In short, the distributed and flexible nature of the message model makes it very difficult to specify—much less capture and preserve—a discrete and bounded record. The complexities increase further when one considers that messages flow and reproduce freely within a complex account and transmission ecosystem.

3.3 Accounts

When a user hits send, disparate systems undertake a complex series of interactions, routing a message to its final destination or destinations. At the most basic level, these interactions rely on accounts and addresses. Creating an email account registers a user with an email provider and establishes a mailbox with an email address, which follows the standard syntax, [username]@[domainname]. Email accounts may represent an individual or a group of individuals. Additionally, an individual may have multiple distinct addresses or aliases associated with a single mailbox; addresses such as help@samplecompany.com can steer messages to the appropriate staff without sharing their personal addresses; distribution or mailing lists are often used so that multiple mailboxes receive mail sent to a single address.

Access to email accounts is controlled by providers and makes use of usernames (generally the email address) and passwords, as defined by RFC 3501 (Crispin 2003). User applications often store credentials locally to allow seamless access via a client, browser, or smartphone application.

3.4 Data Transmission Model

As Prom notes, email at its core is a forward-and-store technology that revolves around message transfer agents (MTAs) and user agents (UAs).

> The delivery of a message requires interaction between one or more MTAs and one or more UAs. Typically, an email server—such as Microsoft Exchange, Postfix, Sendmail, qmail or Lotus Domino—acts as an MTA. Multiple MTAs move an email message from one computer to another until it reaches its final destination. Once a message has been received by the addressed account, the user accesses the message using a UA, such as a Microsoft Outlook client application, a web-based email application or software on handheld devices. User agents provide a method to view, manage, create and forward messages to one or more designated MTAs. In practice, UAs are client applications directly controlled by a user, and MTAs are email server applications indirectly controlled via a UA. (2011, 9)

The BITS Security Working Group in 2013 reported that "The prevalence of phishing attacks has caused a decline in consumer trust of email. In a 2010 Identity Theft Resource Center survey, 81% of consumer respondents cited phishing emails as a significant concern relating to the security of their personal and financial information when conducting online transactions. Email continues to play a role as a significant propagation vector in the spread of malware, causing 54 million U.S. adults in 2011 to report desktop malware infections. The Gartner Group estimates that more than 40% of U.S. consumers have altered their level of trust in email messages and online shopping as a result of this continuing threat" (BITS Security Program 2013, 8).

The forward-and-store nature of transmission affects the archivist's ability to preserve email. Email should be thought of as being in an almost constant state of transition where it is encoded and packaged, decoded and read, decomposed and stored, and reconstituted and exported in a representation format. But what is considered the record copy? The email that was composed and sent? The one that was received and stored in the sent mail folder of the sender's client application? The copy stored on the recipient's email server system? A downloaded copy? A copy stored locally in a PST file? A copy on an iPhone?

Archivists can only capture and store an email message as it exists at a particular point in time and as it is rendered or provided for a particular purpose. The decision to preserve one type or another has important implications.

To date, archivists have most often captured copies of email messages found on a server or device at the end of a period of active use.[14] Commercial email compliance tools, on the other hand, capture a complete version of the email message and envelope at the time it is sent or received. These approaches have strengths and weaknesses. End-of-life capture makes it possible to preserve a record of actions taken on those messages that are found in the device or system (such as opening a message); journaling systems, which copy inbound or outbound emails at the time of transmission to a location outside the email system, where they are protected from deletion or alteration, offer a more complete record of transactions and an audit trail, including messages subsequently deleted by the user.

[14] For a provocative analysis of the implications of this approach, see Bearman 2017.

3.5 Vulnerabilities of Email

Security was not a significant consideration in the original design of email, and many approaches have been taken to grapple with the range of security problems posed by its open and flexible architecture. These solutions are by no means comprehensive or universally adopted and therefore continue to evolve.

Email was designed to allow a variety of human and system actors to modify virtually all aspects of individual messages (body and header fields). As a result, it is vulnerable to many forms of abuse, misuse, or error, including the following:

- **Unsolicited communication,** or spam.
- **Malicious content,** including some targeted at machines (malware, viruses) and some at people (phishing, or plays to steal a user's credentials).
- **Forgery.** There are numerous ways that the authenticity of email content (message body or any header fields) may be compromised. Email "spoofing," where an email is made to look as if it were sent by someone other than the true sender, is perhaps the most significant form of this.
- **Theft.** Because email can be sent in clear text over the public internet, its messages or metadata may be viewed or accessed by unintended recipients who have no authority or rights to view or access the content. Unless countermeasures are imposed, personal or private information can be easily stolen at many points along its transit path.
- **Integrity.** Since all aspects of an email can be altered for valid purposes, there are many ways that email messages can be (intentionally or not) altered or corrupted.
- **Deniability (or repudiation).** Given email's other vulnerabilities, it is plausible for someone to claim that a message was tampered with or even to deny they sent it.

These basic vulnerabilities can be combined in any number of ways. Phishing, for example, is defined as an attempt to obtain sensitive information by disguising a malicious actor as a trustworthy sender. This is a clear breach of authenticity, but sophisticated attacks can breach several vulnerabilities in concert.

3.6 Beyond the ASCII Message: Additional Components

While the defined message structure lies at the core of the email ecosystem, supplementary components add complexity to individual messages, to threads, and to email accounts, turning many of them into an amalgamation of digital objects that pose distinct preservation challenges. Second-order material, such as attached documents and links to external resources, may be as valuable as the message itself, but are easily lost or disassociated from their original context.

3.6.1 Attachments

The original email specification from 1982, RFC 822, stated that email messages must be encoded as ASCII text and placed length and line limits on each message. Users soon wanted to include sound files, images, and documents, and email has since evolved to support the inclusion of such binary objects. To achieve this objective without altering the transmission and network methods, the relevant specifications mandate that binary files must be included within the message body itself and in MIME format.[15] To overcome the restriction limiting email to ASCII text, the MIME specifications require that binary files be encoded with Base64 encoding, which converts nontextual information into ASCII text for transmission and assigns it a MIME type identifier so that it can be decoded and accessed properly upon receipt. The name Base64 indicates that this method of encoding allows the use of only the 64 human readable characters in the ASCII table. Three bytes of binary data are encoded as 6 bits of ASCII data, and upon receipt the Base64 information must be decoded for presentation to the user. Some storage formats, including EML, maintain attachments as Base64 encoded information.

Additional RFC specifications have extended or refined the use of MIME attachments to allow specialized content types such as digital voicemail files and fax transmissions:

- Email as a carrier for voicemail: RFC 3801 (G. M. Vaudreuil and Parsons 2004). Also: RFC 3773, RFC 4239, RFC 6381, RFC 4393.
- Other data types formalized for email:
  - RFC 4142: Full-mode Fax Profile for Internet Mail (FFPIM)
  - RFC 1767: MIME Encapsulation of EDI Objects

Email attachments pose many preservation challenges, not the least of which is that the files themselves pose the same preservation challenges as any other binary data. It's not just the limitless variety of file formats that can be presented as attachments, but also preservation storage of this data and maintaining its relationship to the email message and more. Moreover, as noted in the sidebar, potentially harmful contents can be masked and widely distributed through email.

An archivist's desire to capture attachments is facilitated by the fact that systems allow for their import and export, but complexity arises because emails are created and handled by different clients and servers and because the RFCs do not mandate internal handling mechanisms, storage formats, and export/import capacities. Each system handles attachments differently and sometimes in a proprietary manner. Some clients embed stored attachments in the message itself, in MIME format; others store them separately and in the native binary format. In the latter case, some clients place a pointer in the message; others put the pointer in a proprietary database.

While processing the American Land Alliance collection, including an email account, archivists at the Library of Congress discovered many message attachments with names like "Secret.zip." During the initial appraisal, these messages and attachments were put aside for later processing because their content within the Zip file could not be previewed. It was later discovered that these attachments contained latent viruses. The experience highlights the fact that Zip files and archive files can contain Easter eggs with unknown content and should be opened only with caution.

[15] A complex standard, MIME is broken into five separate IETF RFC documents, each of which has been updated several times: https://tools.ietf.org/html/rfc2045, https://tools.ietf.org/html/rfc2046, https://tools.ietf.org/html/rfc2047, https://tools.ietf.org/html/rfc2048, and https://tools.ietf.org/html/rfc2049.

3.6.2 Links and Resources Outside the Message

If attachments seem challenging, linked resources—that is, content found at a hyperlink and stored outside the email server or client datastore—are even more so. In fact, connections to out-of-message content really make up a separate area of study. Whoever is doing the collecting needs to define the scope of this effort, using the discussions in this section as a starting point.

Some linked data connections act like footnotes. They provide the location (typically a URL) and sometimes the title of a resource. With this information, the reader may be able to find the referenced resource by clicking the hyperlink, but the email client or server does not reproduce the linked content or guarantee access. However, this analogy is imperfect. The networked location is likely to be much more fragile than traditional footnoted items, if only because users are expected to access a resource shortly after receiving the message. A footnoted book, for instance, tends not to be unique and has unchanging content (assuming the correct edition is available). If important enough, a copy can generally be found.

In April 2017, April the Giraffe took the internet by storm as 14 million people watched the livestream of the birth of her calf (Spangler 2017). The link to The Giraffe Cam was widely distributed across the internet, including through email, Facebook, chat, and instant messaging to share the live birth (Animal Adventure Park 2018). As livestreamed video, the link now shows the current activities of the giraffes, not the widely viewed birth.

Other external content is less like a footnote and more like a photograph slipped into a letter. In other words, it functions like an attachment, but with one key difference: the content is included by reference instead of being transmitted through the email system. Simple examples of this phenomenon, such as links to photo galleries or Google documents, illustrate that such resources may not be able to be located after the fact, for a variety of reasons: links can break (there is no longer a valid pointer to the external resources), or the linked material may not be accessible to the archivist (it is stored in a secure or remote location). Even if the linked resource can be found, it may be transitory, as described in the sidebar. Online content such as documents and image galleries show the same malleability, calling into question the authenticity and evidential value of linked resources when email is captured for archival purposes.

3.6.3 Signature Blocks

While they are nominally a component of the message body, email signature blocks often contain essential but unstructured facts about the message sender, including his or her institutional affiliation, contact information, links to social media accounts or other networked resources, and occasionally a logo or other image. For these reasons, signatures can be a trove of information for name variation and associations; in addition, they may supply context for matrix relationships and conversations, provided that repositories can capture the data in a structured fashion and, potentially, resolve it to the URIs for authority records. At the same time, the repetitive data in signature blocks can pose challenges for machine-driven classification processes and skew results.

From relatively simple structured text, email has evolved into a complex record type with significant preservation challenges. Understanding the technical properties of messages (including how they are formatted and transmitted between accounts, the interrelated

standards that define them, and the ways that email applications store and maintain them) is critical for developing policies and systems that will support institutional objectives of preserving and providing access to collections of email.

4. Current Services and Trends

Human and computer interactions around email affect the ability of the cultural heritage community to capture, process, preserve, and provide access to email archives. While the community has partially addressed some of the most obvious challenges of email archiving, archivists and digital preservation professionals must fully grapple with a host of less immediately perceptible issues. One starting point for understanding the current state of play is to monitor and assess email services and trends in the broader IT industry and in society at large.

4.1 The Evolving Email Ecosystem

The significance and ubiquity of email has created markets for technology companies and service providers to create, manage, preserve, and manipulate email content. Whether the topic is email production, delivery, and consumption; integration with other communication services; or legal compliance, many organizations and individuals interact with email systems that spring from a multibillion-dollar IT sector. The platforms that technology companies have developed leverage the open and flexible architecture that made email so popular in the first place, both supporting and complicating the cultural heritage community's mandate to preserve email as a record of past human activity.

4.1.1 Abuse, Abuse Prevention, Security, and Deliverability

Since its early days, email's open, decentralized architecture has created opportunities for abuse, such as forged headers, spam, spoofing, and phishing. As email adoption spread, new security technologies and standards were introduced to block malicious email. This gave rise to a new problem: some legitimate emails were no longer being delivered. Email abuse and crime outpace the security measures developed to stop it. As new security measures are put in place, new "deliverability" practices and technologies arise, and so on.

To work most effectively, email security solutions require systemic adoption.[16] By collaborating, stakeholder and industry groups develop standards that aim to achieve shared goals, while balancing or mitigating conflicts. Relevant groups undertaking such work include the Messaging, Malware and Mobile Anti-Abuse Working Group (M^{3}AAWG), the Email Sender and Provider Coalition, the Online Trust Alliance, and CAUCE, an all-volunteer advocacy

[16] A number of previous standards efforts, such as Author Domain Signing Practices, have failed due to lack of wide adoption (Leiba 2013).

organization.[17] Some important standards are supported by purpose-built organizations, such as Domain Message Authentication Reporting & Conformance (DMARC), which aims to protect against direct domain spoofing (dmarc.org 2018). Many organizations provide a wealth of information that is helpful for understanding email security and operation generally, as well as specifics regarding standards and best practices, including adoption rates and effectiveness assessments.[18]

The prevention of email abuse does have at least one significant negative side effect: hampered deliverability of legitimate email. One report estimates that only 79 percent of commercial (business to consumer) emails are received at the intended inbox because of spam filtering, among other factors (Return Path 2015). Email delivery companies have a vested interest in ensuring that their clients are sending only legitimate email, and many of them are actively involved in the anti-abuse groups.

The community of email services has taken steps such as the following to address security issues:

- **Encryption.** Cryptographic techniques convert information into a code that can be deciphered only with a unique key. Encryption does not prevent data from being captured by malicious actors, but they cannot interpret or use the data without the necessary key.
- **Digital signatures.** Another cryptographic technique, similar to encryption, uses public keys. A signature is produced with a private key to which only the signer has access. The signature can be decrypted using a public key. Digital signatures provide a solid means of authenticating the signer.
- **Reporting.** Numerous security methods use some form of reporting among actors in the email environment. For example, an email provider gives users an option to report particular messages as spam, and when this happens, the sender (or the entire domain) is added to a blacklist so that future messages will be blocked.
- **Content analysis and filtering.** There are many techniques for analyzing the content of email to determine the topic or authenticity of the message, including scanning for viruses, filtering on keywords (e.g., "fast cash"), or applying machine-learning algorithms.

Several security features arising from industry have potential value to the cultural heritage community:

- **Endpoint protection.** The cybersecurity market is enormous and includes secure email gateways and "endpoint protection platforms," which typically include email-relevant features, such as spam and phish deletion (Firstbrook and Wynne 2015; Morgan 2015). These services are largely devoted to real-time detection

[17] See https://www.m3aawg.org/; http://www.espcoalition.org/; https://otalliance.org/; and http://www.cauce.org/about.html.

[18] For example, see https://www.m3aawg.org/supporting-documents and https://www.m3aawg.org/for-the-industry/published-comments.

and prevention but include some features that may be relevant to email archivists: blacklists (domains blocked by email servers) and whitelists (accredited or approved domains). These services provide useful evidence toward confirming or denying an email message's status. For example, the ability to say that a particular email originated from a blacklisted domain would help a user determine its trustworthiness. That said, task force members are not aware of previous work to preserve blacklists or whitelists, much less track changes in these constantly evolving registers.

- **Address verification.** Email security companies also provide services that protect the reputations of their clients as senders of legitimate email by verifying and cleaning their email address lists (see, for example, Informatica 2018; Never Bounce 2018). These services identify illegitimate addresses, such as accounts known to send spam or phishing attacks, so that organizations do not inadvertently interact with corrupt actors (and thus open themselves to attack). Preserving lists of such accounts might help the cultural heritage community. Like blacklists, they could be used after the fact to identify questionable senders within a large email archive. Again, we are not aware of any current work to preserve these lists, much less make them accessible to end users or machine-actionable in a processing workflow.
- **Digital signatures.** Authentication technologies have evolved in a battle to keep one step ahead of spammers. The supporting documents (Task Force on Technical Approaches for Email Archives 2018b) provide a description of the most important security protocols, which include the Sender Policy Framework (SPF), Domain Keys Identified Mail (DKIM), and DMARC, all of which have achieved RFC status. The latter is now widely adopted, and where DMARC signatures exist, it may be possible for archivists or users to make stronger claims of authenticity regarding a particular message. However, this is also an area where the community would benefit from basic and applied research. For example, envelope testing would help us understand whether and how different servers retain envelope information, particularly digital signatures, and whether sufficient information will exist to verify them in the future. Perhaps archivists or users can use DMARC or functions such as envelope journaling to verify authenticity, providing increased confidence that they are documenting the particular transaction that is represented in the message header fields.

4.1.2 Marketing and eCommerce Services

A significant proportion of email is created for marketing and transactional purposes, such as sending receipts and order updates. The technologies and tools used to support marketing and sales continue to evolve, driven by trends such as the increase in mobile device usage, changing consumer behavior, and new products, resulting in an ever-increasing deluge of such emails making their way to the inboxes of often hapless end users. While many cultural heritage institutions will not be interested in such records, others may find reason

to keep them for their evidential value. The underlying technologies also support the distribution of newsletters or other information, which are more likely to be of archival interest.[19]

Email produced with marketing services, customer support systems, and e-commerce applications has some unique features, with implications for email preservation:

- **Responsive email design** builds on responsive web design technologies and techniques to create content suited to the user's email client; the display of mail viewed on a smartphone with a small screen will be different from the display of the same message presented on a desktop computer (Email Design Reference 2018).
- **Real-time email** delivers content that is dynamic, reflecting location, time, or social media interactions. For example, email may include a news, weather, or social media feed that is updated at the moment a user views the email.[20]
- **Video email** brings another layer of technical complexity because most web applications and some clients render video content, and an increasing number of marketing systems allow corporations to embed it.[21]

The challenges of preserving this content are similar to those of preserving any dynamic content, and industry has provided some resources that can help archivists better understand how such systems operate. Several companies provide email marketing guides,[22] as well as the means to track metrics such as client share (users' devices/email clients), deliverability (whether email is received by the intended recipient), and engagement (whether users click through to a website).[23]

4.1.3 Consumer Email Services

Services such as Gmail, Outlook.com, and Yahoo! Mail provide email accounts, software, and services for consumers for personal or small business use. As more consumers go online, the use of these services has increased, because email remains an essential part of the online experience (Radicati Group, Inc. 2016). While its core features and standards have not changed significantly, email services continue to evolve.

---

[19] The email marketing landscape provides an interesting classification of companies providing email marketing technology and services. Gartner provides, for a fee, an annual report on the email marketing sector. Most of the leading vendors require payment to allow the reports to be downloaded (J. Cohen 2015; Hopkins and Sarner 2015; SparkPost 2015).

[20] See https://www.marketingcloud.com/blog/real-time-email-examples/ and http://www.realtime.email/rte-resource/exploring-the-benefits-of-realtime-email-white-paper/.

[21] See https://www.campaignmonitor.com/resources/guides/video-in-email/.

[22] See https://www.campaignmonitor.com/dev-resources/, http://templates.mailchimp.com/, and https://litmus.com/resources for examples.

[23] See https://mailchimp.com/resources/research/email-marketing-benchmarks/ for examples of the types of data that may be provided.

Storage. Email storage limits keep expanding, making many user accounts into illusory archives, which allow seemingly endless storage but dubious preservation. In 2004, Gmail was introduced with a storage limit of 1 GB, more than 250 times the amount then provided for free on Yahoo! Mail, and 500 times as much as the Hotmail service from Microsoft (Pogue 2004). Gmail now provides 15 GB of storage on the free version of the service, and Yahoo! Mail boasts a 1 TB limit.[24]

Filing and search. In 2010, Gmail introduced "labels" as an alternative to the traditional approach of filing emails in folders, claiming significantly improved search capabilities for labeled email left in the inbox (Rodden and Leggett 2010). Most email services now provide similar capabilities, in addition to automatic categorization or sorting of emails, without fully replacing the familiar folder functions (Casserly 2017).

These improvements affect how individuals manage their email and how archivists can process it. Email collections are growing larger, and much of the information that helps users navigate these large collections is not available in the standard metadata supplied by most representation formats, suggesting another area for research and development, since this metadata may provide valuable evidence or discovery pathways.

4.1.4 Enterprise Email Services and Operations

Enterprise email products and services are geared toward organizations that want to operate email for their staff or members, either on their own equipment or on a hosted/cloud application. While consumer email trends largely apply in this area, enterprise vendors have sought to meet market demand by expanding their email products and services to incorporate capabilities such as retention management, compliance, and e-discovery, which are often marked as email-archiving services. Gartner estimated that 10 percent of such compliance-driven archiving was done natively by email platforms in 2016, but predicted that this number would rise to 35 percent by 2021 (Dayley et al. 2016).[25] Institutions archiving email for long-term value may be able to leverage some of these features without additional cost when they are provided or bundled with the core email service. But additional research and development is needed, since these services are targeted not at long-term retention but at short-term management for compliance.

Using such services as part of an effort to preserve email for the long term would require cooperation and support from an institution's IT department or service provider, but the first step is simply knowing what questions to ask. It may be possible for archivists, records managers, and IT professionals in government to assess existing journaling software and to work with vendors to develop

[24] See (Wikipedia 2017b) and https://overview.mail.yahoo.com/, accessed Dec 20, 2017.

[25] Gartner also published a detailed report devoted to evaluating the "native" e-discovery and archiving capabilities of the email software provided by Microsoft (Landers, Harris, and Zhang 2017).

functional requirements for additional system features that would better meet long-term archival needs.

4.1.5 Email Storage, Compliance, and Records Management

There is a significant business need for managing and storing email data outside of email systems. Some industry-supported services focus on storing solely email, while many records management applications manage email alongside other electronic record types. In either case, the existence of external storage systems brings both benefits and risks for those seeking to preserve email for its archival value, as opposed to a short-term business need.

Ongoing regulatory and technology cases, as well as evolving business needs, are reflected in a confusing mix of overlapping terms and categories. The Association of Information and Image Management (AIIM) provides a useful glossary to help understand and distinguish the differences between electronic records management (ERM), document management (DM), and enterprise content management (ECM), among others (AIIM 2018). In regulated industries such as financial services, the concepts are often discussed under the umbrella of "compliance" (Lin 2016). Vendors sometimes use this term to describe features that in other contexts are "records management" or "content management" (Microsoft 2011; Mimecast 2018). Leaving aside semantic differences, the following capabilities are particularly relevant to the archival community.

Storage management was a primary driver for many of the early archiving systems. As email inboxes (not to mention other types of data) grew in size, organizations looked to reduce storage costs by moving some email to secondary (typically cheaper) systems, often at the expense of retrieval speed. Gartner's report on "Enterprise Information Archiving" notes that most vendors started by providing this core capability, often in conjunction with backup and recovery capabilities (Dayley et al. 2016).

Retention management is the ability to specify the minimum period of time that certain records shall be securely stored. This is generally supported by records classification features, where set business rules match records to predefined retention periods or schedules, then ensure that those records are kept for the required period of time before being deleted. In other words, records management systems typically use the term *retention* to refer to the minimum amount of time a record must be kept.

Unfortunately, email systems may define retention in precisely the opposite way: as the *maximum* amount of time records may be kept before being automatically removed from the system. For example, Microsoft Office 365 (which includes email in addition to other applications), supports email retention. But this is simply the greatest length of time an email is to be kept before the system automatically deletes it or moves it to an archive. Before that point in time, it does not prevent a user from deleting it manually; it will just disappear from a user's view and remain in a preservation store (Palarchio 2015). To ensure an email in Outlook 365 is kept for a minimum

period of time and cannot be deleted, use of the "hold" feature is required.

Hold features (also called in-place hold and legal hold) are designed to mark records or emails that must be kept (Microsoft Exchange Online 2017). While holds can be set to expire after specific time periods, they are typically open-ended, and the emails are deleted once litigation is over (or the threat of litigation has passed).

Journaling copies inbound or outbound emails at the time of transmission, based on rules defined by the organization. Typically, journaled copies are stored outside of the email system, where they are protected from deletion or alteration by email users (Microsoft Exchange Online 2016). Another form of journaling, called "envelope journaling," preserves the message envelope in addition to the message itself. This includes information such as Bcc: (blind copying) addresses and distribution list information.

Audit logging or **activities logging** tracks unauthorized access to a user's mailbox and records actions performed by the administrator (or indeed any user) of a system. Some archiving software vendors use this approach to demonstrate data immutability. Audit logging is often considered one of several techniques for maintaining the security of data.

Information rights management (also known as **digital rights management**) enables email users or administrators to use a tool in the email software or a third-party platform to control who can access, forward, print, or copy sensitive email data (Kekre 2015; Microsoft Developer Network 2011).

For archivists, these capabilities present some opportunities. NARA describes journaling and crawling in a user guide for managing NARA email records (NARA 2013b). Journaling captures all email messages and calendar appointments as sent or received. Crawling is a daily process that archives the items and applies automatic records declaration rules as defined by the system administrators. Crawled results are accessible to the mailbox holder, but journaled messages are not. Fitzgerald discusses the use of archiving software to identify, extract, and prepare digital archival records for ingest into a digital repository (Fitzgerald 2013). It is conceivable that audit logs could prove to be highly useful in documenting the provenance of email collections, but the task force is unaware of any such projects. More work should be done to research and investigate how these tools can facilitate archival work.

These capabilities present some challenges. Email stored in external systems may not be stored in an optimal format or include complete metadata. Users may need special instructions or access privileges to extract messages. Information or digital rights management systems may put restrictions on certain types of content (e.g., attachments in emails) that can be changed only by that software (which could be outside the control of archives staff). Overall, the use of information rights management tools has to date focused more on a desire to delete email, treating it as an object of risk management, rather than to preserve it as part of the historical record.

4.1.6 Compliance and Legal Tools

Email is an important record in many legal proceedings and may be entered into evidence under the rules of a controlling jurisdiction. In addition, some case law exists regarding the admissibility of electronically stored information, including email (Pratt, n.d.; Wikipedia 2016). Over the last decade, the legal industry has seen the growth of a substantial market for "e-discovery" services and technology, estimated at $1.8 billion annually (Zhang, Logan, and Landers 2014). E-discovery solutions address several technical challenges related to the long-term preservation of email: controlling large message stores, identifying responsive emails, protecting sensitive information, and preserving email's value as evidence.

Many e-discovery solutions provide journaling technologies that securely collect email (and other documentary evidence such as text or social media messages) at the time of transmission or receipt. Such software can ensure the integrity of files and metadata, and record the chain of custody, but it is much more widely implemented as a means of meeting legal compliance needs than for cultural heritage reasons.

Because email evidence is such an important part of legal discovery, an entire industry supplies the legal community with tools to mine accumulated bodies of email, including those captured with journaling software and those stored on active email servers, client computers, and hard drives. Tools such as PinPoint Lab's Harvester product can directly connect to a variety of email servers, scan and identify email formats from hard drives, copy data without altering the original file metadata, and prove integrity using hash verification (Pinpoint Labs 2018). Self-collection kits guide data custodians through the collection process and automatically record all actions to document the chain of custody.

E-discovery solutions use a range of text mining and analytics techniques to help human users find specific information within a body of text. Because the machine and person operate in a feedback loop, such applications are commonly called technology-assisted review or TAR.[26] While many industries use sophisticated text mining technology, the legal industry has very high standards for effectiveness, which must be defensible in court: the purpose of TAR is often to find records that will help a prosecutor or attorney tell a story, or make a case, regarding past events (Attfield and Chapin 2018).

The alternative is to manually review documents, but given the increasing volume of electronic documents, this process becomes very costly and time-consuming. Accordingly, the legal industry has developed a significant body of knowledge on how and when to use these technologies effectively, and the community has released large email data sets, which can be used for training and testing purposes (The Coalition of Technology Resources for Lawyers 2016; Duke Law Center for Judicial Studies 2018). A leading provider of e-discovery

[26] The terms *auto-categorization* and *predictive coding*, meaning the use of keyword search, filtering, and sampling, are also used to describe semi-automated portions of an e-discovery document review (Exterro 2018).

software has even published a case study on how to apply its technology to identify and remove sensitive data and personally identifiable information from an email corpus (Nuix 2018). Gartner provides annual reports on the state of the e-discovery/TAR technology market and vendors (Zhang, Logan, and Landers 2014).

E-discovery techniques are not without controversy and are subject to significant study and discussion (Grossman and Cormack 2014). However, studies have shown that human review of documents for sensitive, confidential, or privileged information can be prone to error. In some cases, TAR and auto-categorization have been shown to be more effective (Cormack and Grossman 2017; Grossman and Cormack 2011). If only because the results of their machine-learning algorithms are rather opaque, we cannot directly infer that these technologies will meet the archival community's standards for identifying sensitive or personally identifiable information. Nor can we assume the opposite: that they will never meet archival standards. While the costs of these tools may put them beyond the reach of most cultural heritage institutions, the University of Illinois is currently leading a project to assess their potential usability in the context of government and academic archives (Illinois State Archives, and Records and Information Management Services, University of Illinois at Urbana-Champaign 2017).

4.2 Challenges for Repositories

Alongside work being completed by industry, the archives and library community is developing its own set of approaches to acquiring, processing, preserving, and providing access to email. In considering the range of options that are now available, most archivists will face a set of specific questions and decision points, which can be grouped roughly into the following areas:

1. What email should my repository capture, and how?
2. How can my repository maintain email in a way that facilitates its value as evidence?
3. How can staff who are processing email adhere to archival actions for acquiring, appraising, processing, and preserving email?
4. What specific techniques can be used to deal with attachments and linked content?
5. How can security and privacy issues be mitigated?
6. What special challenges are introduced when working with large collections, or with many collections?

While much progress has been made in answering these questions, the following narrative also notes areas where the completion of additional work would benefit the community as a whole.

4.2.1 Capturing Email

Given the varied nature of the originating systems and sources that might supply email of archival value, the first task in any repository workflow—capturing email accounts and messages as objects that

are amenable to archival process—is a complex one. Accordingly, the cultural heritage sector has not, to date, established best practices for capturing email from existing systems into a standardized, preservation-ready form.

For organizations, a culture of institutional risk avoidance poses the first and most prominent barrier. Reputation-conscious organizations sense that the liabilities of preserving email far outweigh the potential benefits, placing archivists in a defensive posture. For example, legal counsel may argue that since destruction is allowed, email should be deleted so that it cannot be subpoenaed.

That said, technical barriers exacerbate the cultural challenges. To access email mailboxes it may be necessary to use a variety of often poorly documented applications capable of imposing difficult-to-remove constraints and dependencies. Given the array of platforms that people use to access email, archivists may not know which instance to capture. Server, desktop, web-based, and mobile applications all require different export methods and may supply different formats or metadata. With the range of email mailbox and message formats available, there is often a need to convert the messages to a format that is compatible with a given tool set. This conversion is not always lossless and can itself lead to significant metadata loss. Good donor and IT relationships help, but they depend on the particular technical skills and access that such partners bring to the table and are not always productive.

Attachments and email threads pose another set of capture problems. It is not enough just to capture an attachment: the relation to its parent message must also be maintained to preserve the context that makes both parts valuable. This task is complicated by the diverse ways in which attachments can be stored and associated with individual messages; it may not become apparent that attachments have become disassociated until well after capture. In the same vein, the thread of replies in email mailboxes should be maintained to preserve the original context of the correspondence.

Staff and time limits pose additional constraints in two ways. First, with limited access to a donor's or department's email account, archival staff must make quick decisions. This often results in a mixed bag of content, with valuable records being kept alongside materials clearly out of scope. One strategy to deal with this issue—to push appraisal to the donor—is fraught with complications. In theory, donors might do more careful appraisal, but many lack the time, inclination, or knowledge to follow through. Donor-initiated appraisal can lead to the donation of the entire mailbox, only a small subset, a scrubbed group of records, or no email at all. Also, there are few (if any) tools across email systems that make it easy for a records creator to do selection without "cleanup" of the inbox. A partnership between archivist and donor seems like a much better model.

Email archives and data files can be found, at times unexpectedly, on physical media such as hard drives or removable media. Files may have been placed there as a result of proactive backups or may be found in system libraries hidden to all but the most tech-savvy

donors. These email archives are often not discovered until years after accession or active use when a digital backlog is processed, precluding appraisal or discussion with the donor at the time of capture.

To deal with such issues, some institutions have developed local guidelines for capturing email collections. More often, repositories have captured email on an ad hoc basis, with tools assembled and chained into just-in-time workflows. The duplication of effort with institutions focused on hyperlocal efforts slows the momentum across the profession as a whole, leading to a deprioritization of email capture because of its perceived difficulties. Left unresolved, this situation could become a vicious cycle with the problems and complexity increasing the longer the issue remains unaddressed.

Current practice points toward some promising signs of convergence, with four capture scenarios currently predominant. While they all illustrate the impact of the aforementioned issues, some common themes emerge:

- **Direct export.** When possible, direct export is a preferred method of capturing email, but it typically requires working with IT staff and is more common for institutional records than for donated collections. Some tools (such as ePADD and Preservica) support direct Internet Message Access Protocol (IMAP) connections. In this scenario, the archivist can work with IT staff to ensure that the export format is compatible with the institution's tool set and that the metadata exported reflects the email's significant properties. There is also a potential for high-level appraisal before capture, particularly if the archivist can define particular export rules for implementation by IT staff.
- **Web service export.** The archivist or donor exports email from the web-based application and sends it to the repository. Compared with direct export, this method is a bit more prone to format and conversion problems, as well as metadata loss. Attachments and email threads may also be captured in a way that undermines their context. However, this method does allow the archivist some potential for high-level or detailed appraisal before capture.
- **Client-based exports.** Microsoft Outlook supports the export of an entire account or a portion of it as a PST file; Apple's Mail program supports the export in MBOX; and Google has a "Takeout" feature that allows email export. However, data files may have to be converted to another format, and the profession currently lacks tools to confirm lossless exports or correct attachment migration.
- **Disk imaging.** If archivists have direct access to physical media that contain email, the email can be exported from a disk image of that media. Since the archivist is not making the primary capture, there is no control over formats, metadata, attachments, and email threads. It is likely that they will be working with an internal, and potentially proprietary, data store. There is no potential for appraisal before capture, and interaction with the donor is likely to be limited.

In spite of the seeming differences in these approaches, exports tend to employ a few semi-standardized formats, such as MBOX, EML, and PST. Each tool may or may not include particular metadata. As Prom reports, it is likely that many migration tools preserve well-defined header and message body properties (as defined in The InSpect Project [Investigating the Significant Properties of Electronic Content Over Time]) (Prom 2011, 10; Grace, Knight, and Montague 2009). However, little work has been done to specifically test how well particular tools do or to follow up on the report's recommendation that communities define significant property profiles that might apply to particular content areas. As a simple example, it would be helpful to know whether particular tools include information such as whether an email was read and if they preserve metadata such as flags or keywords (as defined in RFC 5322). Furthermore, the significant properties could be brought forward to treat actively incorporated extensions. For example, RFC 5451 added an "authentication-results" header, which is now in wide use and of potential importance in making claims about the authenticity of particular messages (Kucherawy 2009). Moreover, local or custom header fields could be meaningful. Take the case of institutions (particularly in government) where voicemail is captured and sent through email (Vaudreuil and Parsons 2004). Many of these systems use proprietary header fields, for example, to record the phone number of the person who left the message.

It should also be noted that each of the four approaches work at the end of the active use of email and assume that email will be captured from stored copies. This has two major implications.

First, it results in a much attenuated email record. Waiting until an email account is no longer used may result in fewer messages because of deletion by the creator or by the application of retention rules imposed by legal or system storage limit policies. When email is captured for an inactive account, the chances of data loss increase. System configuration variables may be missing, and digital signatures may no longer be valid. Rolling capture (e.g., capturing select email on an ongoing basis) may enable more content to be acquired and integrated with other materials, but it may also create technical challenges around deduplication and message threading. Journaling software allows real-time capture of all traffic. Rolling accessions (such as annual PST captures from active accounts) may risk continuity of the account itself, resulting in complicated differential and deduplication exercises, or they might be missing components if the process is not executed precisely during each capture. For these reasons, the method and frequency of transfer will impact long-term preservation goals and complicate efforts to document provenance.

Second, the community must address the fact that capturing email at the end of its active use imposes steep trade-offs and does not take advantage of industry-standard approaches for capturing email at point of transmission. As suggested by attendees at a Council of State Archivists/National Historical Publication and Records Commission symposium on email archives (papers not yet publicly

available), the government records community would benefit greatly from a project to define functional specifications and modifications to journaling tools that would better facilitate the identification, real-time capture, and long-term management of messages known to or likely to have archival value. Such a project would be particularly valuable if undertaken in collaboration with state chief information officers and representatives of major enterprise email service providers, such as Google and Microsoft.

4.2.2 Ensuring Authenticity

If a digital object such as email can be shown to be what it purports to be, we say that it is authentic. Authenticity, according to Rothenberg in the CLIR report, *Authenticity in a Digital Environment*, "is intended to include issues of integrity, completeness, correctness, validity, faithfulness to an original, meaningfulness, and suitability for an intended purpose" (Rothenberg 2000, 52).

Given the primacy that email standards place on openness versus security, it is not surprising that email is susceptible to forgery, modification, deletion, and decontextualization. In the legal realm, questions surrounding email authenticity date to the 1990s (Bearman 2017). *Armstrong v. Executive Office of the President* stated that the electronic copy is the primary copy and that the paper printouts of email were convenience copies, incomplete because they lacked structural and contextual information and metadata such as headers, links, and timestamps.

For email collections, authenticity might be formally demonstrated through a series of documented processes and events. For example, if a business email was created and maintained in the course of normal business using an approved email application, if it contains the official email address and key elements of the message (e.g., header information, attachments, signature blocks), and if it is preserved as part of recordkeeping processes under the terms of a records retention schedule and in a system that prevents unauthorized modification, it is a reasonable assumption that the message is authentic. Unfortunately, things are not so clear in the real world, and attempts to judge authenticity from a forensic perspective are arduous (Banday 2011). Similarly, email system documentation and server logs could be valuable data points for authenticity validation, as could certificates used to authenticate a message upon receipt. Registration and electronic document and records management system (EDRMS) environments also create natural processes and infrastructure for maintaining authenticity, but few archives use such records management systems, at least not directly.

Ultimately, authenticity is judged by a person assessing evidence and forming perceptions. A recent InterPARES study concluded that people use contextual information and functional characteristics of email as primary criteria for determining authenticity and that—quite significantly—material preserved in established archives or library is accorded a high presumption of authenticity (Bunn et al. 2015).

Web archives provide interesting parallels when considering provenance and authenticity. Scholars, including Ankerson (2012), Ben-David and Huurdeman (2014), and Maemura, Becker, and Milligan (2016), have shown that the lack of explicit provenance metadata for web archives (e.g., why, how, and when web content is captured) is a challenge for quality scholarship. The research ethos provides an underlying rationale for better stewardship. The key takeaways are that researchers who use large-scale, born-networked archival corpora—such as email and web archives—are struggling to determine how to provide context and how to provide a steady state from which other researchers can verify the validity of their findings. While researchers obviously play a role in documenting their computational and analytical methods, the archives should likewise provide access to provenance metadata as well as to preservation and discovery system requirements and parameters.

4.2.3 Tracking Processing and Preservation Actions

While an archive's staff may have relatively little control over external factors impacting message authenticity, they can take specific actions to demonstrate the archive's trustworthiness. For a typical digital repository, actions such as the following show considerable good faith in documenting a collection's provenance, ensuring its chain of custody, and tracking its processing history:

- Registering the transfer of ownership or custody of the material
- Ensuring contextual information is retained; for instance, repositories should preserve attributes such as the folder structure of an email account, the relationship between emails and their associated attachments, the relationship between the email account and any other digital archive material being transferred, and additional metadata that might exist about the material
- Maintaining a full audit trail of any actions taken on the material, and the person or system responsible for carrying these out
- Running fixity checks on the material when it is copied or moved from one storage location or medium to another
- Recording repository actions as part of the preservation metadata that accompanies the email throughout its life

As this list implies, the principle of provenance can be applied to email without significant deviation from that applied to other digital formats. That said, most archivists lack access to a full-blown, integrated toolset to support active, ongoing documentation about the actions they perform on email corpora. Institutions without automated systems can manually record information such as that shown in the model on the following page, then store it electronically in their collection management database.

For some or all of these actions, a large institution may have access to a mature preservation repository infrastructure that can record the results of automated processes. Ideally, these records will become part of the digital provenance and help system administrators track digital objects for long-term preservation and inform

Archivist-created Provenance and Processing Metadata Model

Context of email creation

- Systems used to create email (e.g., web-based Gmail or Microsoft Outlook client)
- Platforms used (computing environment)
- Type of use (business, personal, joint email account with family or organization, extracurricular)
- Who used the account (e.g., person or organizational unit comprised of persons)

Context of use or recordkeeping

- Account details
- Account name/username and legal/official name
- Length of time using the account
- System of arrangement (e.g., folders present or not and how they were used, whether the trash folder was used for drafts or for review)

Context of preservation or curation (many of these activities can be documented as PREMIS events)

- Who received the materials
- What acquisition processes were used to acquire the content
- What tools were used to inspect, appraise, inventory, review, and describe the materials
- Preservation policies that have been applied to the materials over time (when, by whom, for what purpose)

For personal papers or the donated records of an external organization, additional metadata may need to be tracked, such as:

- Selection criteria
- Email systems used by the donor
- Method by which an archivist captured the records

user judgments regarding authenticity.[27] Appendix A provides an example of the automated capture of metadata, as implemented by Harvard University.

Ideally, technology will be used to integrate documentation from the digital preservation system with the collections management system (or vice versa). Such integration would function to create, produce, and compile a unified and holistic dataset of information about what has been done to the objects over time. Assigning tracking identifiers for digital objects and associating activities with them will be crucial if an integrated system is developed.

The fact that such tools are not more widely available points to an opportunity: the community could enhance modern collection management software to better integrate the capture of information regarding actions taken on email collections. Ideally, data would be

[27] PREMIS is the community-accepted standard for preservation metadata. The *PREMIS Data Dictionary* defines a core set of semantic units for the preservation functions of digital repositories and the data model defines entities. For an excellent synopsis of the use of PREMIS, see Caplan 2009.

captured throughout the archival workflow, from the point of selection through appraisal, accession, arrangement, description, and preservation. Some systems already include digital object modules where an archivist can link events or activities to single messages, groups of messages, or even entire accounts.

When archives use digital preservation systems, the captured metadata usually documents preservation actions such as file format identification, migration, and ingest. These records are automatically collected as system-generated actions on the backend in the form of PREMIS records or some equivalent. However, other metadata that is not typically recorded by digital preservation tracking systems could provide end users with critical information. For example, records describing appraisal decisions, description activities, and the results of privacy evaluations would help end users judge the authenticity and context of a collection. For repositories that lack connectivity between their collection management system and digital preservation repository (or for archivists who don't have digital preservation software), this documentation could take the form of processing notes, similar to Light and Hyry's notion of using annotations and colophons as a new approach to commenting out the work contained within a finding aid (2002). The sidebar provides one example of how such data could be tracked in a human-readable format.

On October 9, 2017, Jane Doe, processing archivist at the University of Delta, used BitCurator to review John Doe's email account in the John Doe Papers Collection to identify any materials that were out of scope per the donation agreement. The tool scanned the email account based on donor-provided search terms and found 35 emails that were out of scope per the donor's wishes. Jane extracted them from the email account and recreated an MBOX file of the email account with the extracted items removed. Jane re-ran a file format identification tool called JHOVE to confirm that the new MBOX was properly formed. File format was determined valid, so then she moved the MBOX file into the staging area where another archivist would start additional processing work later in October. All of these actions were recorded manually by Jane, as a processing note in the University of Delta's collection management system.

4.2.4 Preserving Attachments and Linked Content

Attachments pose particular challenges for the archivist processing an email collection, beginning with the seemingly simple tasks of ensuring that an attachment remains linked to its original message and that an archivist and user can render it for viewing. Less immediately obvious but even more problematic is content that is linked to a message through a URL. The URL might be visible to the end user, or it may be embedded in a way that makes it appear to be an integral part of the message—although it may actually be stored on a server distant from the main body of the email message.

Archivists must address the following areas when working with attachments and linked content.

Potential for Loss or Corruption During Email Conversion. Throughout the email workflow, conversion tools may not properly handle attachments, resulting in complete loss or corruption from an incorrect conversion. When capturing emails and preparing them for processing, it is often necessary to convert them from proprietary email client formats to a standard format such as EML or MBOX. Similarly, certain client or server applications may export emails in a way that does not fully preserve attachments. This issue can be manually addressed by monitoring the results of the conversion through the application of quality control procedures. The time- and resource-intensive nature of such work suggests that semi-automated testing regimens would better facilitate reproducible workflows within and across repositories.

A 2016 NARA research project determined than many common virus-checking software tools could not detect viruses within PST files. In fact, NARA could not determine that any tool could successfully perform this task. NARA determined that the best course of action is to reformat PST files into individual EML files and run the virus checker on the EML files and associated attachments.

Processing Issues. When working with email collections, archivists and curators should be cognizant of several risks. By carefully attending to the following factors, they can ensure that the preserved files are more authentic and useful to future researchers:

- **Potential deletion:** It is easy to accidentally delete an entire email when only the attachment is to be preserved. For embedded attachments, the attachment must be extracted from the email in order to be preserved separately from the email message. This suggests that the processing history of attachments should be tracked in addition to the history of the message itself. Most critically, systems should record the removal of attachments or disassociation of attachments from the message. Automated removal of attachments especially comes into play when replying to or forwarding a message, so archivists should also be careful when deduplicating collections or threads
- **Format issues:** Attachments may be of many different formats, each potentially requiring a different tool for identification and evaluation. However, technical and structural review of attachments may be accomplished with commercial software such as Quick View Plus or FTK (which contains Quick View Plus).
- **File size:** Some attachments, such as video, may be very large, placing a strain on storage systems and, where files are transferred across networks, on the network. Archivists should also be aware that some systems store large files externally at the time of mailing or receipt. Gmail, for example, can be configured to automatically store large objects in Google drive and create a link in the message, while Apple's Mail Drop provides similar functionality (Gmail Help 2018; Apple 2018). Both systems introduce access and retention restrictions that might render attachments inaccessible in the future. Preservation is a constant risk with email systems, which routinely add new features to enhance the user's system experience.
- **Viruses and malicious content:** Attachments may contain viruses or other malicious content. These must be detected and procedures put in place for determining how to handle such files.

Repository/Preservation Issues. Once an email account or set of messages has been processed, an archival repository must confront storage and long-term preservation issues, focusing on two themes.

- **Attachment storage and handling:** When an attachment is embedded within the email message in MIME format, it is relatively easy to guarantee that the attachment will remain associated with that email. However, attachments are easiest to monitor and store in their native binary format, not as MIME content embedded in a message. To ensure that the stored file can be located, a pointer should be placed in the original message. Many email client applications do this automatically, if somewhat invisibly. In the context of a digital repository, which must preserve content for long periods of time, there are many unresolved questions about how to maintain fidelity between the message and attachment. But the

upshot is that repository software should have a method for the persistent linking of attachments and email messages. At the simplest level, the relationship between a uniquely identified message and attachment could be documented in a spreadsheet or METS file. A more complex solution might entail modifying the body of the message to insert a pointer to the new file location. Any solution will require careful planning, as the storage location may be moved at any point in the processing workflow or during subsequent maintenance of the repository.

- **Migration:** Files attached to emails take on entirely new format preservation issues over time when compared with the body of an email message, which is typically just ASCII or UNICODE text. This suggests that preservation policies based on email formats (for example, deciding on a target storage format for the message itself) will not adequately address the preservation of attachments over time. If attachments are placed in a repository, they can presumably be migrated to formats that constitute a target preservation format. But this raises a set of additional questions: Will the original attachment be retained? How will the target file be associated with the message? Will formats be monitored over time, and if so, how will additional migrations be tracked?

Linked Content. Preserving linked content in email collections is a parallel challenge to preserving native content within the message or attachment. Many organizations now require or request that staff save files in file management systems (e.g., SharePoint or Box) and only provide a link to the intended content via email. In this case, email acts as the connection to the data but does not contain the data itself. While this is inherently a business decision—helpful for saving space and avoiding data collisions when many people are to approve or review the same document—it does have significant implications for archiving. In terms of benefits, the risk of exposure is lowered if the link leads to a Box folder requiring authentication. If the email message is inadvertently forwarded or the email account is hacked, the unintended recipient can't access the file. Yet security is a double-edged sword. Without the record creator's credentials or access to the folder, the archivist or user can't access the file either, perhaps long after any need for confidentiality has passed. For these reasons, the archivist may work with the donor to acquire a separate copy of the files, obviating any need to rely on an external system.

In other institutions, the decision to preserve linked content might be more flexible, informed by available resources and collection development policy. Since the linked content is stored outside the message, archivists may decide not to preserve it. Such a strategy has, of course, precedent in the analog world; archivists would not necessarily track down a publication or other document that is mentioned in a folder of correspondence, but would simply assume that interested users could try to track down such documents as part of the research process. In some cases, however, archivists may decide that it is important to save such content, based on collection analysis or some other factor.

The presence of such linked content is increasingly problematic when integrating email with file management and messaging systems such as Sharepoint, Slack, Box, Dropbox, and Google Drive. The linked content may, in fact, be integral to—if not more important than—the message itself. As anyone who has tried to access a photograph or document while off the network knows, such content is simply inaccessible and will not render in the message if it cannot be located. With networked resources, these issues become more prominent over time: locations can change, content can change, or the content might simply never have been preserved. Referenced network resources may not be accessible if they are located behind a firewall, on an internal network, or in another location that can be accessed only by authenticated users. Some links have been run through a URL-shortening service/link management platform such as bit.ly. Because these services produce a URL that may have no association with the actual location, even the most diligent researcher would be unable to definitively identify the original resource from the URL. Some, like bit.ly, promise to maintain the link permanently, but this will always be contingent on the continued existence of the service.

Email-archiving applications have not, as yet, provided functionality to harvest resources at a URL that might be found in a message, much less to preserve embedded images, documents, or other content. A solution to mitigate the second situation would be for the application to recognize known services and follow the link to record the actual location of the resource. To do so, the application would have to be kept up-to-date with major services and would provide only an original location for the resource. Another, and more comprehensive, solution would be to link the email-archiving application to software that could retrieve the resource or store it at an allied location, such as a web archive collection or as a linked resource (similar to an attachment). For the sake of practicality, this would have to happen in a timely manner—such as in a message journaling workflow—to capture the linked content before changes occur. The key issue is whether tools used in email archiving should extract web links, attempt to crawl those pages, mint a DOI or some other persistent ID for the archived Web page/site, then replace the link in the archived email with the new permalink. This question highlights the challenges of both context and scale for email archives.

Adding the linked content to the email collection would ensure that it continued to exist, though at the price of increased storage cost and complexity. Because it is often impossible to know the full extent of a web resource, web archiving scope could become an issue as well. Finally, it is hard to imagine what solution can be devised for situations in which the location is inaccessible because the resource has been moved or deleted, or exists within a private network, other than allowing a harvester to authenticate to that location with the provided credentials.

In short, accounting for attachments and linked resources in email collections is a difficult task, but one for which institutions and developers will have to account. The set of issues identified previously suggests many potential avenues for development.

4.2.5 Ensuring Security and Privacy

Security and privacy are core concerns with active accounts, and such issues do not disappear once a collection arrives in an archival repository. On the contrary, an archives must take specific steps to ensure that email collections remain secure and that private or otherwise sensitive information is not compromised.

Personal, Sensitive, Restricted, and Classified Information. Perplexing issues surround email collections that include information subject to temporary or permanent access limits. In a government or organization, some messages may be formally classified as restricted or privileged and confidential. Other messages are not marked as such, but also require sensitive handling or include information that must be restricted by law or policy, such as student records, health information, or personnel records. Often, few if any such messages are classified or marked for restriction; they accumulate with other traffic on an account and can be difficult to identify, much less segregate.

Because of its integration in our personal and professional lives, email often includes private information of a more personal nature. The names of family members, telephone numbers, and other contact information is often included, as might discussions of a sensitive nature around health and wellness. Other messages are intended for a specific audience with need-to-know only access. These issues affect messages not only in work accounts, but even more so in personal collections that might be donated to an archives. When third parties are involved (e.g., people mentioned in messages who have no knowledge that information concerning them is being donated to a library or archives), these issues can seem intractable.

Issues regarding private, sensitive, and restricted content will impact most aspects of stewarding email accounts from initial acquisition through access and preservation. Management will be governed by copyright laws, legal agreements, or contracts with owners, as well as other public laws (Whitt 2017). Given that most email accounts include a mix of personal, transactional, and professional interactions, archivists should use secure methods for acquiring, working with, and accessing these accounts. This need for security will affect and shape workflows, policies, and tools as archivists ingest, appraise, process, deliver, and preserve these accounts. How and how much private data is acquired and preserved is often the result of negotiations between donors or email account owners, records management guidelines, policies of the institutions, federal and local laws, and redaction and access systems. The task force has provided a more detailed examination of email security standards, available on the Task Force on Technical Approaches for Email Archives project website (2018b).

Encryption. Another type of security issue must be considered: the capability for encryption at the email message, folder, or account level. Formats with strong functionality for encryption (such as Lotus Notes, in which "email messages can be encrypted before sending,

after receipt or before saving") are widely adopted in certain industries where privacy is key (Sustainability of Digital Formats 2015). Formats such as Outlook's PST also support encryption, but it may not be consistently implemented (Sustainability of Digital Formats 2013a, 2013b). Encryption, as a rule, is antithetical to digital preservation unless the encryption key is provided. There is currently no clear solution to these issues, other than the archivist gaining access to the encryption key at the time of donation, then saving the materials in an unencrypted format.

4.2.6 Processing High-Volume or Numerous Collections

The scale of email archives varies a great deal among academic and government institutions. Smaller institutions may struggle with appraising, ingesting, and preserving even a single account. An average email archive in an academic setting might contain 40,000 messages, while an archive of one million messages would likely be considered quite large. In federal government institutions, one million messages might be considered small.

Since email collections tend to be so large, and since many, if not most, will require some level of sensitivity review, manual review processes are not fit to the task. Estimates for reviewing Supreme Court Justice Elena Kagan's email indicated that 6,000 work hours would be needed to make the 75,000 "pages" available (NARA 2010). The iterative process of reviewing email from the Virginia Governor's office—to make 150,000 messages available from a pool of 1.3 million—demonstrated similar challenges (Library of Virginia 2016).

NARA developed the Capstone Approach as one method to deal with the issue of scale (NARA 2015). Federal agencies have interpreted the Capstone guidelines in different ways—some identifying offices or individuals of importance, others choosing to preserve all outgoing or incoming messages (Plante 2015). Some repositories and museums have created guidelines for staff to tag or folder emails concerning specific projects or exhibits, where the information needs permanent retention (Rockefeller Archive Center 2006).

But even within these whole-account approaches, the scale of messages, entities, and subjects within email accounts is tremendous. One large email archive at Stanford Libraries contains 700,000 messages. Within it, 385,000 messages were unread, 78,000 were marked as important, 203,000 were in the sent folder, and the remainder had been put into specific folders by the creator. The University of Manchester, likewise, was working on an archive with 170,000 messages (16 GB) in the Carcanet Press Email Preservation Project (Baker 2014). Each account contains tens of thousands of named entities. Even a relatively small email archive, for example, the Robert Creeley email archive at Stanford, consisting of 50,000 messages, contains 24,000 correspondents, 57,000 persons, 42,000 organizations, and 22,000 locations extracted through natural language processing. It also contains more than 10,000 attachments in a variety of formats. The complexity of email archives and the scope of attachment formats, numbers of correspondents, and named entities is daunting.

Impact of Scale on Digital Preservation Processing. Issues related to very large-scale collections, such as those routinely encountered with email, touch every part of the digital preservation lifecycle. Take, for example, appraisal. It simply takes more time for personal review of large-scale collections to determine messages of enduring value. Archivists processing the Carcanet email collection noted significant problems with spam and junk mail; hundreds of genuine messages received from bona fide correspondents were identified as spam in the subject line (Baker 2014).

Email collections, perhaps more than other types of collections of digital content, include high levels of duplication. By the very nature of a transaction between a sender and a receiver, there are at least two copies of an email message created in the digital ether. Add to these messages those sent from listservs or large mailing lists, those received as copies of messages sent to someone else, or those received through a web of "reply all" messages. In addition, people routinely include existing message threads in responses, even when the subject or topic has shifted. For these reasons, there can be many versions of the same messages in a single account or across many accounts.

In an ideal world, records management guidelines would help people manage institutional email mailboxes to remove duplicative or non-record content. But such guidance is all too often ignored, and many duplicated messages will make it through appraisal and into a processing workflow. Automated processes and tools to help identify true duplicates—not just items that are similar in content, authorship, or language—are essential to facilitate appraisal. Most deduplication tools, such as Treesize Pro, focus on network file management and cannot deduplicate email collections.

The bulkiness of email collections raises concerns about arrangement and description as well. Long-standing archival principles and practices suggest that archivists will and should describe email collections in the aggregate rather than at the item level. This approach is in line with the move toward "More Product, Less Process" across the archival community, but email processing could take better advantage of software tools that can augment the archivist's review of the collection.

A research project at Georgia Tech's Information Technology and Telecommunications Laboratory tested format identification for a variety of document types through a UNIX file command, but its reliability and accuracy are not yet fully demonstrated.[28] The project staff are also developing a method to automatically identify the

[28] The Georgia Tech report says, "A database system for managing file format information and creating the magic file used by the file command is described. The metadata for file formats including file signature tests can be easily changed in the database rather than in the magic file. This is a substantial improvement in the flexibility of the UNIX file command and magic file. A graphical user interface has been developed for the file command. File signature tests have been created for more than 800 file formats and the reliably [sic] of the file command and file signature file is being evaluated on examples of the file formats it purportedly identifies." (Underwood et al. 2009).

topics of e-records, including email. This could facilitate automatic description of the records and help ease subsequent access requests to the collections, such as freedom of information requests.

At the same time, archivists can take advantage of the affordances of digital content such as email to automate procedures and use descriptive methodologies that would be highly challenging for large-scale paper collections. For example, ePADD performs name resolution of correspondents to aggregate aliases and email addresses for individual correspondents (with the ability to edit), in addition to deduplicating messages. The program can also extract entities (persons, corporations, places) that are mentioned in and presumably subjects of the archive. The TOMES (Transforming Online Mail with Embedded Semantics) project is developing a similar feature for government email (NHPRC 2018). These named entities have several uses. By reviewing frequency counts, an archivist can identify some of the most prominent subjects discussed in the collection, then add those as access points to a record describing the collection as a whole. Entities can be compared against local and international authorities, allowing links to national or international authority records.

For access and reference services, email collections need to be human-readable, because researchers will be looking at individual emails in the future. That said, access is moving away from providing snippets to offering corpus-level discovery (as exemplified by ePADD), with full text display, attachment previews, access to header information, and more. Researchers are also working with email as a form of data, with the work of the History Lab and Virginia Governor's Office email as examples.[29]

Need for More Resources with Larger Scale. The large scale of email collections relative to other types of digital content means that more resources are required to preserve it. Dealing with large volumes of data is costly from an infrastructure perspective in terms of storage, computing power, and bandwidth requirements. Ingesting large collections takes computer resources, fast internet connections, and patience. One task force member who used the Google Takeout feature waited two days before receiving a link to download her email archives.

Beyond infrastructure, the second main cost is labor—the time and skills needed to process the data throughout its lifecycle, from donors to archivists to, ultimately, researchers (Rouse 2018). A report regarding eDiscovery conducted by the Rand Corporation Institute for Civil Justice reports that "the application of predictive coding would have saved an estimated 86 percent in attorney review hours" over manual review (Pace and Zakaras 2012, 68). EDiscovery would have cut the eyes-on, hands-on time from 686 hours to 98 hours to review 47,650 sample documents. This is still a significant amount of human-interaction time. It indicates that meaningful review can take

[29] See http://www.history-lab.org/; http://www.virginiamemory.com/collections/kaine/

place only with a commensurate investment in both technologies and the staff to use them.

Researchers at the University of Illinois are examining the use of predictive coding ("the use of keyword search, filtering and sampling to automate portions of an e-discovery document review") to assist in automating electronic discovery, or e-discovery, for processing email records submitted via NARA's Capstone Approach (Illinois State Archives and Records and Information Management Services, University of Illinois at Urbana-Champaign 2017). E-discovery, the "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," can be a complex and burdensome process, made even more so by the sheer volume of data to get through (Rouse 2018). A Rand paper reports that e-discovery mean costs can run approximately $22,480 per gigabyte reviewed (Pace and Zakaras 2012, 28). While large-scale data processing is expensive, solutions such as automated predictive coding can be made more effective over time to reduce costs (Pace and Zakaras 2012). Without a significant advance in such technologies and their full integration into email-processing work, it seems possible, if not likely, that large sections of the historical record will remain closed indefinitely to research, whether that is in support of historical scholarship, documenting rights, or ensuring accountability and effective services.

While automation through predictive coding and the like can bring costs down, money and time aren't the only taxed resources. The community at large must look at using new techniques and approaches, and these require new skills, new technologies, and changes to established working practices, funding, and governance. In short, email archiving requires more than just software or the money to buy it.

5. Potential Solutions and Sample Workflows

While issues persist, tools and the community of practice have developed to the point that institutions can make headway in acquiring, preserving, and providing access to email collections. The trick lies in choosing wisely from a range of preservation approaches and tools to create a workflow that meets local needs. This section of the report provides recommendations for both approaches and workflows. Appendix B provides detailed information about some useful tools and services available to archivists; a more complete list is available on the Task Force on Technical Approaches for Email Archives project website (2018e).

5.1 Preservation Strategies

Archival repositories have a range of preservation options to consider, including bit-storage, migration, and emulation. Preservation planning should embrace the entire lifecycle, but in most circumstances archival repositories will take the lead. They not only

develop plans, but also monitor changes that impact file sustainability and access.

Regardless of the chosen approach, tools should be able to permit the exchange of data about both the email and relevant preservation actions, perhaps using their native APIs or extensions added to existing tools. For institutions dealing with their own organizational email and attachments, there is likely to be greater control and uniformity over the formats created and preserved, based on institutional policy. Collecting institutions taking in email archives from external bodies and individuals will have much more diversity in their holdings and will therefore have to invest more resources into preservation planning.

5.1.1 Bit-Level Preservation

Bit-level preservation is a set of methods and services that protect content over time from threats such as bit-rot and unintended deletion or changes. Typical actions include those that support Level One in the National Digital Stewardship Alliance (NDSA) Levels of Preservation recommendations such as checksum creation and fixity checks, the creation of archival backups, and file format characterization (National Digital Stewardship Alliance 2013). Bit-level preservation does not take into account display, context, and interpretation of the digital object. But it does allow the digital objects to be ingested into a monitored system or digital repository until the need arises to provide active access to researchers and other users.

Bit-level preservation is a valid option for institutional digital repositories that have not yet developed or incorporated functional preservation processes for email collections. It would protect email messages and attachments until they could be fully rendered and made functional.[30] In the meantime, copies of the account or portions of it can be provided manually to users of the archives, provided that messages are managed under an access policy meeting legal and donor requirements. Although the approach is simple, it comes with a higher risk; because the data is preserved in its most basic form, there is no promise or expectation of comparable ease of use after it has been ingested.

5.1.2 Migration

Migration strategies for digital preservation follow two basic pathways, sometimes alone, sometimes in combination:

- Transfer of digital data from a less stable to a more stable format
- Migration of digital objects from the great multiplicity of formats used to create digital materials to a smaller, more manageable number of standard formats that can still encode the complexity of structure and form of the original

It is the latter strategy that most often comes into play for email collections because many of the existing tools and workflows are

[30] Because of the technology dependence of email within proprietary systems, there is often an element of unintentional migration when extracting the email from the account platform before ingest into a repository.

anchored around certain file formats. Rothenberg supports this approach in his paper:

> The need to refresh digital information by copying it onto new media (and possibly translating it into new formats, sometimes called "migration") has been recognized in the library sciences and archives literature, as well as in a number of scientific and commercial fields. This requires an ongoing effort: Future access depends on an unbroken chain of migrations with a cycle time short enough to prevent media from becoming physically unreadable or obsolete before they are copied (Rothenberg 1999, 11).

Format migration is the process of transforming a file from one format to another, usually to mitigate the risk of format obsolescence, but also for the purposes of software tool compatibility or format normalization for file management. It is perhaps the most commonly used preservation strategy to date. It is important to note that migration for email collections typically refers to the message format only and not that of attachments.

Why is migration such a popular path for email? A big reason is that popular tools dictate specific formats. The community has established the MBOX (for email accounts) and EML (for individual messages) specifications as de facto preservation formats because they are reasonably well-documented, nonproprietary, open, and readable and writable from a wide variety of email tools, including ePADD and Harvard's Electronic Archiving System (EAS) (Prom 2011). ePADD ingests only MBOX or IMAP accounts and suggests external programs for migrating to MBOX. In the case of EAS, it ingests many formats but normalizes to EML for processing and preservation.[31]

XML is another format that some institutions have adopted for email messages. As a software and hardware-independent markup language, it is an ideal format for preservation—although there are only limited tools available to render XML-encoded email accounts for access purposes. There is no internationally recognized XML schema for email, but the Smithsonian Institution and the State Archives of North Carolina have developed the Email Account XML Schema (EAXS), which has been adopted by several other U.S.-based archives. As its name suggests, EAXS produces a single XML document holding an entire email account, but it allows several configuration options, including the ability to separate attachments from messages and to save them to an external data store.

---

[31] Harvard plans to release EAS as an open source application in the coming years, but at the time this report was being finalized, use was restricted to Harvard only. See *Digital Preservation Handbook,* "File Formats and Standards" section for more rationale for normalization: "Arguably, in some sectors, proliferation is more of a challenge than obsolescence. If formats aren't normalised then an organization can end up with a large number of different file formats, and versions of those formats: e.g. lots of different versions of PDF, word, image formats etc. In domains which develop rapidly evolving bespoke data formats this problem can be exacerbated. Tracking and managing all these formats—which ones are at risk, and which tools can be used for each one—can be a serious challenge." (Digital Preservation Coalition 2015).

Migration is, at its heart, change, and change always carries with it elements of risk. There is the possibility of data loss during the transformation for any number of reasons, including format incompatibility (i.e., when the destination format cannot accurately carry or identify all the information from the source file), flipped or discarded bits, file corruption, and more. As with any digital preservation strategy, an institution must weigh risk against the potential for reward based on test results, available toolsets, knowledge, skills, and experience (Waters and Garrett 1996).

If employing migration as a preservation strategy, an institution should always retain the email in its native format as well; this essentially forms the original manifestation, and if future developments in technology lead to more effective preservation methods, it will still be possible to revisit the original bitstream and work directly on that.

5.1.3 Emulation

Emulation may be considered an appropriate and useful approach for preserving email in some contexts, particularly where it is considered important for the user of the archive to experience the email inbox of a significant individual in its original context and to immerse themselves in that person's working environment (Loftus 2010).

Emulators are software applications that simulate one set of computing hardware on another set of computing hardware (Daintith and Wright 2008; Wikipedia 2017a). Software is written for specific hardware architectures and can be rendered obsolete if any changes are introduced that prevent its execution. To maintain access to any such software or any digital content that requires that software for interaction, emulators can be used to simulate the older machines on which the software was used. Email is particularly distinctive among digital objects in its relationship to software, since there are multiple points of exchange and hand-off between software applications along the pipeline of processes in which users and administrators interact with email (The Document Foundation Wiki 2018). In cases where archivists choose to provide emulated access to an email collection, metadata describing specific aspects of the computing environment (i.e., email client/application version) will be necessary to emulate the software environment.

While there are many different applications that will enable users to interact with email (e.g., read, annotate, delete, send), and while email standards are relatively simple (compared with those of many other types of digital objects), the software used with born-digital files can change both the experience of that interaction and, more significantly, the content presented to users, such as image attachments or HTML-encoded information (for examples, see Archives New Zealand 2018).

Emulation thus presents a mechanism for users of email collections to interact with the email in the original intended software applications. (It also provides a reasonably certain method to render and view attachments, assuming they were renderable in the original

operating system, and if that entire operating system is emulated.) To enable emulation strategies over the long term, the email applications must be preserved alongside the email to ensure that the content presented to future researchers is not distorted through the use of inappropriate software. In addition, future users may need to engage with the legacy email software to understand such things as the limitations of search functionality, the ease of use of the "reply all" function, and the like.

Email Access in Context. There may be scenarios in which the desktop environment that was used to receive (and create) archived email can be acquired or captured at the same time as the email archives. Using emulation to maintain the ability to interact with those desktop environments ensures that any future researchers will experience the creator's email with the additional context of user settings and preferences. In the case of private email archives, digital forensics tools can be used to take an image of the email archive owner's computer hard drive; the disk image can then be migrated to make it run (and continue to run) on emulated hardware. In the case of email coming from corporate or government sources, it may be possible to acquire copies of the standard desktop environment images that were used on employees' computers, ensuring the necessary software is captured and available for access via emulators in the future.

Emulation Challenges. Like all preservation approaches, emulation has inherent challenges, including the availability and preservation of legacy software as well as the legal protections, especially those governing intellectual property, relevant to software and computer code. In addition, because no single institution has the capacity to collect all the software titles that may be necessary for emulated access to software-dependent collection material such as email attachments, the future of streamlined and equitable access to software-dependent cultural heritage materials calls for a coordinated but distributed international effort to identify, collect, describe, and preserve software. More information about technical approaches to emulation and development of emulation projects is available on the project website for the Task Force on Technical Approaches for Email Archives (2018a).

5.2 Interoperability to Support Flexible Workflow Design

Each of the preservation approaches depends upon combining systems and tools. In short, they rely upon interoperability. The *Oxford English Dictionary Online* defines interoperability as "the ability of two or more computer systems or pieces of software to exchange and subsequently make use of data."

5.2.1 Processing Functionality Across Multiple Tools

In an interoperable environment, archival processing of email may involve the transfer of data (such as accounts, messages, or headers) between multiple systems. Processing email is a progressively iterative activity. Revisiting processing may occur as part of a planned and phased approach, based on the availability of resources, or it may occur when new information about a collection or object comes to light. In addition, technological advances may offer new and enhanced ways to process data, making it worthwhile to revisit an already processed collection. Tools, therefore, may need to access and interpret data at any point in email's lifecycle—including well into the future—so that archivists can view or edit the content (e.g., delete email messages or redact sensitive data) or metadata (e.g., update rights information when an embargo period ends or record preservation actions such as migrating the format of the content), or both.

Content and metadata that can be accessed but not parsed, analyzed, or understood by a tool can still be preserved. If different email tools model metadata differently—for example, one tool might maintain metadata for each individual email message while another does so for a whole account—they can still render or display or associate the metadata with related content and transfer metadata between them. If a tool does not recognize the content or metadata sufficiently to process and deliver it, it can still preserve the data and pass it on to other tools via an exchange package.

These and other issues related to interoperability were discussed at a workshop at Harvard Library in March 2016, which brought together leading practitioners in the field of email archives to discuss the need for tools that manage the full stewardship lifecycle of email and to identify future directions for collaborative work (Harvard Library 2016; Murray and Engle 2015). Participants concluded that any single system or workflow solution was impractical because solutions at each institution would need to correspond to their local policies and objectives. The group discussed a more practical approach focused on the potential to build an environment where tools could be used flexibly and interchangeably—depending on local needs—to support multiple workflows. The ability to mix and match tools increases flexibility within an organization as technology and requirements change, and also facilitates cross-institutional work.

For this type of flexible workflow design to work, the community of practitioners and tool builders needs to agree upon a minimum set of requirements that must be met to exchange and make use of email data (content and metadata). The more the community agrees on what needs to happen, the more functionality will be available across different tools. Therefore, a balanced and flexible approach to building interoperable tools should include functional requirements for common needs defined by the community along with requirements for handling local needs.

An initial look at the inputs and outputs of some of the existing tools began at the Harvard workshop as a way to assess the potential for interoperability. After the workshop, Harvard Library engaged

Artefactual Systems, Inc. to continue collecting data about the tools and to report on the challenges and opportunities for improving their interoperability. The resulting report, *Email Archiving Systems Interoperability*, suggests a set of generic requirements for the interoperability of tools that may provide a starting point for future community exploration (Simpson 2016).

The TOMES project is also working to develop methodologies to move archival email accounts out of proprietary hosted systems. Based on early demonstrations, task force members believe that TOMES will be very effective in helping institutional archives implement Capstone-style acquisition and processing of institutional archives. If TOMES project staff members secure additional grant funding, they hope to develop predictive coding features and recruit libraries to assist in processing those email accounts harvested from proprietary systems (Simpson 2016).

5.2.2 Developing a Community Data Model

The generic requirements in the Harvard report include the need to agree on basic formats and structures for both metadata and content. One way to do this is through data modeling.

A shared abstract data model for archival email would serve as a ground truth resource, helping to future-proof the model against changes that might shift requirements for later work on format and metadata definitions. A community data model would also simplify the process of supporting multiple expressions of the package structure—for example, using METS or JSON. The model could serve as the core to which the appropriate archival or digital preservation requirements can be added. Most tools used for processing email in archives will be able to recognize and use the RFC standards for email, so they seem like a logical starting point for the community data model.[32]

A data model specific to email would include the community-defined structure for an exchange package, the content, the metadata, and the relationships between them. Data models also specify which components and elements are required and which are optional. To increase interoperability, the community could select a specific package format that would organize the content, pointers to content, or both for transfer and exchange between tools. Tools to process email data can be built to leverage the predictability that a data model offers—including the location and type of metadata and content that should be expected—thereby maximizing the tool's functionality.

One example on which the community could build is the content model that Harvard Library established for ingest and storage of email in its preservation repository (Harvard Wiki 2018, 57).

[32] Recommended by Stephen Abrams, associate director of the UC Curation Center (UC3) at the California Digital Library (CDL) in a conversation with Wendy Gogel on October 31, 2017.

5.2.3 Defining Format Requirements

One area where the community can take direct action is in defining the specific formats in which email archives should be stored and exchanged. While some work has begun in this area, additional harmonization would set the stage for better tool integrations.

Email Messages and Attachments. As mentioned earlier in this report, community agreement exists around the MBOX and EML format families as de facto standard formats for data storage and exchange of email messages (Library of Congress 2016; Murray 2014; Prom 2011). Since many tools already include support for input and output of MBOX, it is a good first candidate as a minimum requirement when building tools. Tools that convert from EML to MBOX (and vice versa) permit even more flexibility.

Attachments are a secondary form of content and can be any MIME type. They can therefore be processed, preserved, and accessed by MIME type, or they can be passed along between tools as binary files, with preservation assured by a fixity check and identification through an external identifier.

Metadata. Practitioners working with email might consider implementing PREMIS in METS as one option for modeling a metadata profile, based on the wide adoption of both by the archives and digital preservation communities.

The *PREMIS Data Dictionary for Preservation of Metadata* is considered an accepted international standard (PREMIS Editorial Committee 2008). It defines the core metadata needed by digital repositories to maintain the "availability, identity, persistence, renderability, understandability and authenticity of digital objects over long periods of time" (Lavoie and Gartner 2013, 2). Using PREMIS still requires implementation decisions. The most widely used metadata container format for implementing PREMIS is METS, which can also be used to wrap or point to metadata in other formats (Zierau and Peyrard 2016).

One benefit of using PREMIS is that the metadata can be modeled and defined to support functions that are commonly recognized as useful to the archives and digital preservation communities. There are five "entities" defined in the PREMIS data model, including, for example, those for events and rights (Caplan 2009; Dappert et al. 2013). The events entity is used to record information about changes made to the content during processing, such as format conversion or deletion. The rights entity is used to record rights information, much of which is designed to be actionable. It is easy to imagine how useful it would be to have rights information about the email content recognized across multiple tools. For example, it could be expressed, maintained, and acted upon by multiple tools (manually or automatically) to impose access restrictions.

While PREMIS and METS help facilitate the digital preservation of all kinds of content, the community can achieve even greater functionality in an interoperable environment by also agreeing on an email-specific metadata profile (Caplan 2009; Zierau and Peyrard 2016).

Implementing PREMIS in METS is one option for modeling a metadata profile.

Exchange. Agreeing upon an exchange package format will facilitate transfer between the interoperable tools. Each tool in a workflow should recognize and accept an exchange package; however, they don't all have to be able to unwrap and access the package contents. Intermediary tools could do this and transform the contents to formats accepted by subsequent tools in the processing chain.

These exchange packages that are output by one tool and input by another correspond to the Dissemination Information Package (DIP) and Submission Information Package (SIP) of the Open Archival Information System (OAIS) reference model (OAIS 2012). Many tools produce and unwrap the container format BagIt, developed by the Library of Congress and the California Digital Library, making it a good exchange package format candidate for email (Library of Congress 2018).

5.2.4 APIs and Interoperability

An application programming interface (API) is a protocol in which software tools are configured to represent the content and metadata in a common way at the point of data exchange.

APIs can facilitate interoperability in many ways. Primary among them is their ability to integrate disparate systems without the components needing to know anything about underlying application language and functionality. Without these constraints, system environments can shift and change flexibly and iteratively because they are not locked into specific dependencies, making the microservices approach a real possibility. As needs change, APIs help support unanticipated uses by allowing for the sunsetting of applications as they reach end of life and the development of new tools.

From a functional perspective, APIs can accommodate different data serializations and formats (i.e., XML, HTML, JSON), and they have the ability to leverage several layers or types of security (e.g., transport layer security, encryption). In the web environment, most APIs operate using the principles of Representational State Transfer (REST).

A notable benefit of APIs is that, when thoughtfully designed, they can shield developers from technical complexity. Users don't need to know programming languages for all the different systems they want to integrate, just the language in which API calls are being made. Archivists, particularly at smaller institutions, need tools that don't demand extensive technical skills or infrastructure but still ensure security, integrity, and authenticity. REST APIs may be one tool to help meet such demands. They typically provide an endpoint, or URL prefix, to which a set of parameters may be submitted, specifying the data to return.

Email-Specific REST APIs. Some industry email providers such as Google and Microsoft have made REST APIs available for their products (Google 2017; Microsoft Developer Network 2018). Although these seem primarily geared toward supporting email interactions in applications that are not full-fledged email clients, they may offer some possibilities for email archives. Third parties also offer REST APIs, allowing integration across email providers (Context.io 2018; FWD:Everyone 2018; Nylas 2018).

Potential uses

- APIs allow for further abstraction of interactions with email; it is not necessary to use a specific email client to access email, which facilitates community-developed API libraries.
- APIs could serve as a backup or verification of other email export or transfer tools. For example, a user could ping the API to get a count of messages and make sure that is what resulted after exporting email with Google Takeout.
- Generally, these APIs seem to support more real-time and granular access to information. This opens the possibility for rolling or incremental transfer of email over a period of time rather than a one-time account dump. While it is unclear if that need currently exists in the cultural heritage sector, it is an untapped feature that has high potential for use by organizational and corporate archives.

Cons

- APIs do not support the transfer of an entire account via a single endpoint. Transferring an entire account would require several API calls (at least) and likely many more (especially for Gmail, where it may be necessary to make a call for each individual message).
- Many API responses produce individual messages in a JSON response rather than in openly documented and widely supported formats such as MBOX or Maildir.[33] In most cases, the RFC 2822–formatted email appears to be available within the response, but additional analysis should be done to assess whether all header information is provided.
- A potential solution to these challenges may be JSON Mail Application Protocol, or JMAP, an emerging JSON-based standard supporting communication between email stores and client applications that is designed to standardize data structures as well as make more efficient use of network resources. The IETF JMAP Working Group is actively working on draft documents describing the data model and protocol with a new version released in early May 2018. This project is conducted in the GitHub repository associated with the proposed standard.

[33] Maildir is an email format in which each message is stored in a separate file with a unique name, so it isn't affected by operations on other messages. Created by Daniel Bernstein for the qmail MTA but now implemented in other programs, Maildir is designed to "to eliminate program code having to handle locking" (Maildir 2018).

Potential Barriers and Risks to API Integration. APIs pose some potential implementation barriers for email processing. Notably, most email processing takes place at the account level, requiring the completion of lengthy processes on large numbers of messages before the next tool can take over. This blocks easy implementation of a microservices approach to email processing, where one tool might hand off a stream of objects to successor processes, while other objects are being processed in a prior stage. In addition, the following considerations must be addressed if APIs are to become more widely used in email-archiving applications:

- Not every application has an API, so integration is not always possible.
- Poor or incomplete documentation can undermine the usefulness of APIs.
- An API-based ecosystem may force developers to learn API endpoints and interaction patterns instead of bringing detailed local knowledge of applications to bear.
- Scalability, particularly for bulk operations, may be a problem because not all services can handle large operations.
- Large data volumes that eat up bandwidth may create additional costs or complexity in throttling or error detection.
- APIs provide a level of abstraction from system functionality. This abstraction may not always be useful and, in some cases, may hinder troubleshooting.

API Workflow Scenarios. There are two areas in which publicly documented APIs may be useful in the archival management of email if a common exchange format could be developed as a prerequisite to additional tool development:

1. Providing connections between lifecycle phases. For example, tools that support the appraisal of email should provide APIs to allow interaction with tools that support accessioning and archival processing. Although most tools have robust import/export functionality (some of which is built around email APIs such as IMAP and SMTP) many of them lack more general purpose APIs. In addition, email processing and discovery tools could export metadata to external descriptive systems. In other words, tools developed by the cultural heritage community, as well as externally developed tools, could use APIs such as JMAP to foster closer workflow integrations.
2. Performing bulk operations. Because virtually all archival functions with email records happen in bulk, tools that perform bulk operations on email will also benefit from publicly documented APIs, since these operations can often be automated to increase efficiencies. This would include, for example, topic modeling for appraisal, format migration, and fixity checking.

5.3 Workflows and Implementation Scenarios

This section of the report introduces common processing workflows. While not intended to be prescriptive, they offer examples of the range of practices that are currently supported and point to areas where greater interoperability would improve the processing experience and outcomes.

5.3.1 Bit-Level Preservation Workflow Scenario

Creator Relations and Pre-acquisition Appraisal

Ms. Sanchez, a writer of regional significance, wishes to donate her research papers, including electronic records, to a small local historical society. The historical society has mostly paper-based records but is starting to accept more and more digital material; it is eager to build its capabilities as resources permit. It has a fledgling digital collections effort with one digital-savvy archivist, but to date, most of the society's focus has been on preserving and making word-processing documents and pictorial material available to users. While it does not yet have the capacity to make the email collection that Ms. Sanchez wishes to include with the donation available for research, the historical society wants to accept it, hoping for better preservation and access options in the future.

The donor agreement specifies that although Ms. Sanchez has several email accounts, she is donating only the one she uses for professional research—a Microsoft Exchange Outlook account through an Office 365 personal subscription—to the historical society and keeping her personal Gmail accounts for her own continued use. The archivist from the historical society, Michelle, interviews Ms. Sanchez about the content and context of the email account and learns that this account is largely devoid of personally identifiable information and other personally sensitive information, so little pre-appraisal is needed.

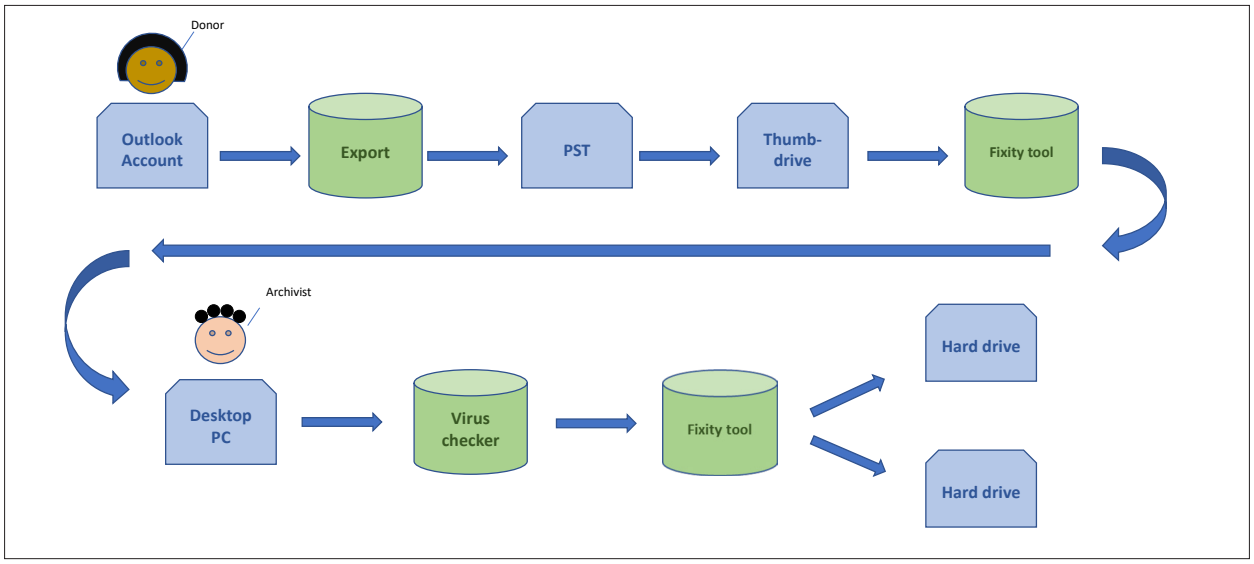

Fig. 3: Bit-level preservation basic workflow

Capturing Email

Following Michelle's advice, Ms. Sanchez uses the export functionality of Microsoft Outlook to export the contents of the account to a PST file, which she adds to the thumb drive containing all her digital records destined for the historical society (figure 3). The historical society also instructed her to use the Fixity open-source tool to establish a checksum for each file on the thumb drive (AV Preserve 2018).

Appraisal, Processing, and Storage

Michelle accepts the thumb drive containing the data, including the PST file, and copies the data to a local drive on her office PC. She runs a virus checker followed by the Fixity tool to check for authenticity issues and, finding none, copies the data to two external hard drives for storage—one kept in the collections storage room, and one in the processing area. These hard drives offer some redundancy until such point as they can be ingested into a preservation repository.

5.3.2 Migration Workflow Scenarios

Migration Workflow Scenario 1: Harvard Library

Donor Relations and Pre-acquisition Appraisal

1. Keith, digital archivist at Baker Library Special Collections, Harvard Business School, receives a hard drive from a faculty donor—Dr. F—as she is preparing to retire.
2. Keith uses FTK Imager to create a disk image of the drive and then uses FTK to investigate the contents. He unexpectedly discovers a PST file containing Dr. F's Outlook email. Using FTK, Keith extracts the PST file.
3. Keith contacts Dr. F, who was unaware that her email was on the hard drive and is hesitant to agree to donate it. Keith offers her the opportunity to review the email first, reiterating how important her correspondence is to her legacy.
4. When she agrees to review the email, she mentions that she no longer has that account, as she closed it out the previous month. Keith offers her space in the reading room. He is planning to set up a machine running the ePADD Appraisal module for her. To use ePADD, he needs to convert the PST file to MBOX, so he requests conversion in a secure environment from the Library IT department. As the staff has an Emailchemy license, they use the tool to convert the PST file to MBOX.
5. Keith imports the MBOX file into the ePADD Appraisal module to prepare for Dr. F's review.
6. When Dr. F visits the reading room, Keith teaches her how to search her email using the ePADD Appraisal module and to flag what she would like to donate to the collections at the Business School. She is pleased that she can identify and exclude the correspondence between herself, her husband, and her son and is convinced that the rest of the correspondence will enhance the value of her faculty archives. When Dr. F is happy with the final results

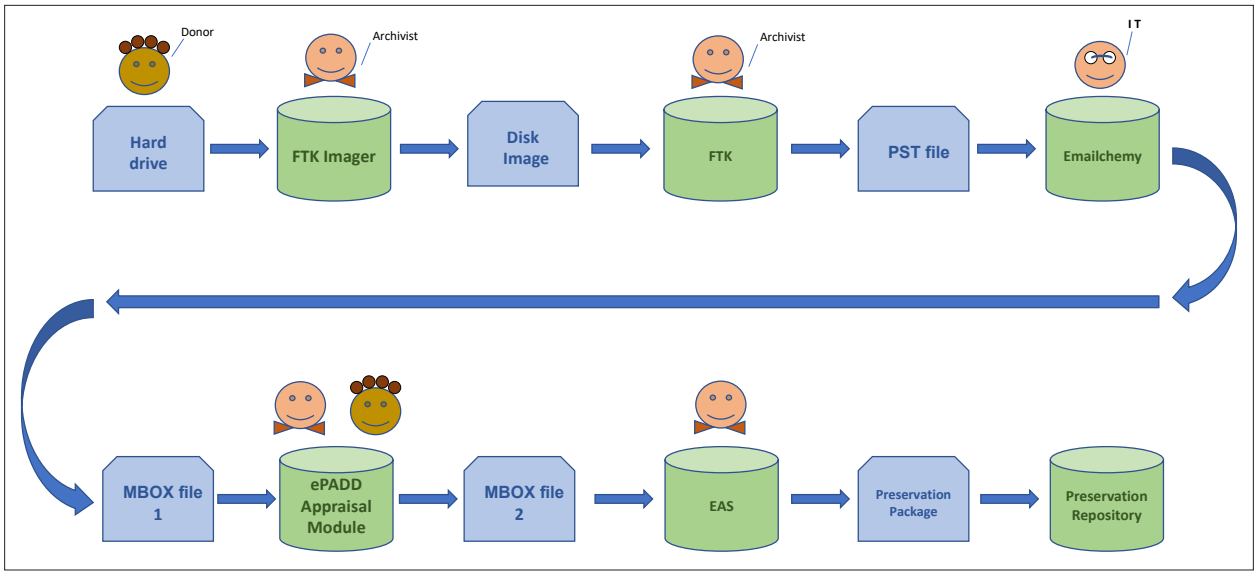

Fig. 4: Harvard Library migration workflow scenario

of her appraisal, Keith exports the MBOX file for acquisition by Baker Library Special Collections (figure 4).

Accession, Processing, and Deposit to a Preservation Repository

7. As an archivist at Harvard Library, Keith can use the EAS for processing email and depositing it to the university's preservation repository. He drops the MBOX file exported from ePADD into a secure drop box for EAS. In EAS, Keith applies a small amount of metadata, including accession ID and collection association. As part of the loading process, EAS uses Emailchemy to convert the email messages to EML and applies the metadata to each email message and attachment. Since he knows that Dr. F served as an advisor to a future U.S. president, he devises a set of search criteria to identify relevant email and applies a series title to them (to help those processing the collection in the future). Keith also devises searches to identify student, personnel, and administrative records to apply embargo periods according to university guidelines and to flag several of the email messages and attachments as requiring restricted access. He then pushes a button which automatically packages the content (email messages and attachments) and metadata and deposits it for long-term storage in Harvard's preservation repository. Keith is not addressing discovery and delivery at this time since all of the email messages and attachments are under embargo, in accordance with the university's policy for donated faculty papers.

Migration Workflow Scenario 2: Stanford Libraries

Creator Relations and Pre-acquisition Appraisal

ePADD is intended to support the creator of an email archive who has intimate knowledge of the contents, by enabling them to review and flag sensitive materials before the email is transferred to a

repository. As an example of how that process would work, consider the case of Mr. and Mrs. Hache—who have decided to donate their collection to Stanford Libraries' Department of Special Collections (figure 5).

After an agreement has been signed, the project archivist or curator—possibly in tandem with a digital archivist—visits the creators of the collection. The repository team interviews the couple about their digital content, workflow, and the like. They discover that, among the many formats in the collection, the couple has email—in fact, three email accounts with different online clients.

The Haches have: (1) a Gmail account, which is still active; (2) a defunct AOL account; and (3) an older Outlook account, which is no longer used and is stored on an external hard drive.

Capturing Email

1. For the Gmail account, Mrs. Hache uses Google Takeout to select and download the account. Google will notify her when an MBOX file of the account is available for download. She downloads the MBOX file 1 and sends it to Freya at Stanford via FTP.
2. At Stanford, Freya imports Mr. and Mrs. Hache's old AOL email account directly from the AOL IMAP server, using a login provided by Mrs. Hache, in the ePADD Appraisal module.
3. For the defunct Outlook messages, Peter, digital archivist at Stanford, uses FTK Imager in Stanford's born-digital lab to create a disk image of the drive and then uses FTK to investigate the contents. He discovers a PST file containing Outlook email, extracts it using FTK, then converts it to an MBOX file 2 using Aid4Mail.

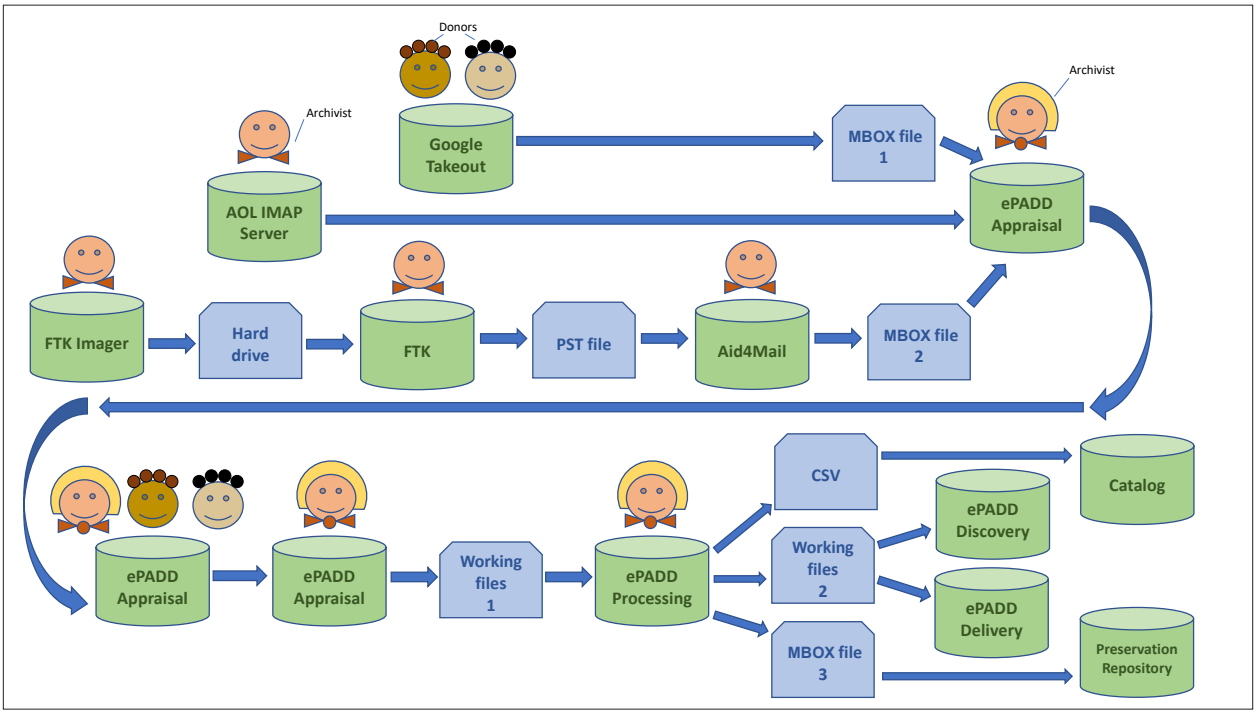

Fig. 5: Stanford Libraries migration workflow scenario

Appraisal

4. In preparation for Mr. and Mrs. Hache's visit, Freya imports the MBOX files 1 and 2 into an instance of the ePADD Appraisal module installed at Stanford Libraries, the same one where the AOL email account was loaded.
5. When the Haches arrive, Freya teaches them how to search their email and instructs them on how to flag messages and attachments to exclude them from transfer or flag them with an embargo period.

   After quickly identifying personal correspondence among their family members and messages relating to medical issues, the Haches exclude them from transfer to the repository. While there is functionality available in ePADD to identify and flag sensitive messages (which can include regular expression and lexicon searches for personally identifiable information, entity identification for diseases and syndromes, and image attachment review), the Haches do not take advantage of all of it. They are satisfied that they have identified the most personal material and allow the rest to be included with their papers.

Processing

6. After Mr. and Mrs. Hache leave, Freya exports all of the email messages except those marked "do not transfer" by the Haches from ePADD's Appraisal module. The export produces working files 1, which Freya imports into ePADD's Processing module.
7. In the Processing module, Freya:
   a. Exports an MBOX file 3 (preservation copy). This file contains the email that the Haches are donating and will be ingested into a digital preservation repository.
   b. Runs further processing tasks on the emails to prepare for research access. This includes:
      i. identifying and excluding sensitive, restricted, or legally protected content and exporting processed email messages as working files 2 (access copy); and
      ii. engaging in data cleanup, authority work, and selection of names for export to a CSV file for import into a catalog record.

Accessioning, Preservation, and Access

8. Freya imports the CSV file into the catalog record at Stanford (and a finding aid where applicable).
9. When a researcher requests access to the collection, Freya will import the working files 2 (access copy) into the ePADD Discovery and Delivery modules.
   a. In the Discovery module, a standalone web application at Stanford, researchers can remotely browse and search a redacted email collection prior to physically traveling to a repository's reading room to access the full corpus.
   b. Using the Delivery module at a managed workstation in a reading room at Stanford, the researcher accesses the full contents of the unrestricted portions of the access copy, including attachments.

10. Freya deposits MBOX file 3 (preservation copy) into a digital preservation repository.

Migration Workflow Scenario 3: Smithsonian Institution Archives

Appraisal and Accessioning

1. Head of exhibits at Smithsonian Institution Carl is leaving the institution, and his email correspondence is identified for appraisal and accessioning by acquisition archivist Jennifer.
2. The Smithsonian's electronic records archivist Lynda submits a formal request for access to the IT department. She works with the email system administrator in IT to capture Carl's active and archived email accounts.
3. Lynda uses MessageSave to normalize the captured email files from Carl's two accounts to MBOX format, producing MBOX files (set 1).
4. Lynda loads the resulting MBOX files (set 1) into DArcMail, which automatically generates metadata.

Processing, Preservation, and Deposit to a Preservation Repository

5. In DArcMail, Jennifer and Lynda work together to process the email accounts:
   a. Jennifer reviews the metadata and determines that processing is required before the accession can be finalized.
   b. Jennifer and Lynda analyze the contents and devise search queries using layered Boolean techniques across email fields, body text, and message attachment filenames to aid in weeding and in selecting the final sets of email for accession.

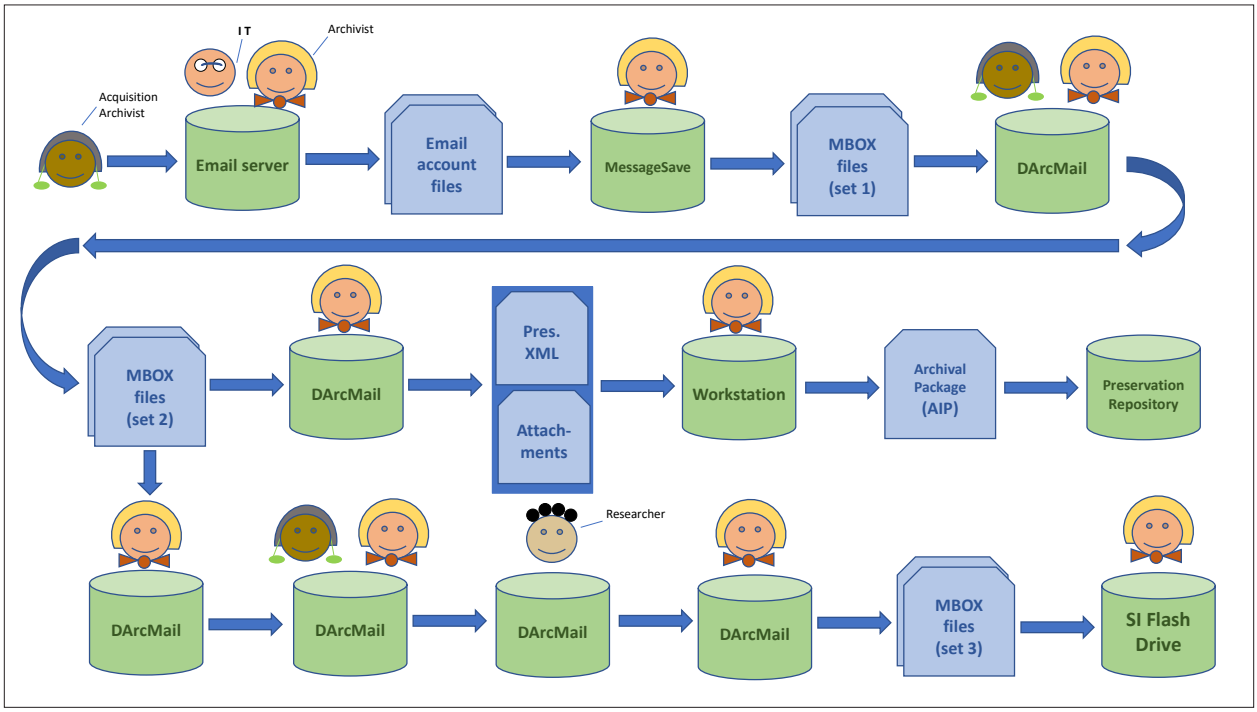

Fig. 6: Smithsonian Institution Archives XML migration workflow scenario

6. The selected final accession sets of email are exported from DArcMail as a second, reviewed and weeded, set of MBOX files (set 2).
7. Using DArcMail, Lynda selects the MBOX files that have been accessioned and migrates them to their preservation XML format with DArcMail. As part of this step, DArcMail
   a. creates fixity values for each message in the set and for the set itself;
   b. generates metadata for the accessioned content; and
   c. separates email attachments into a parallel directory structure.
8. On her local workstation, Lynda creates an archival package (AIP) comprising the original email account files, the normalized MBOX files (set 1), and the DArcMail output (the preservation XML and email attachments) in a folder. The AIP folder is transferred via SFTP to the preservation repository for long-term retention.

Access

9. Authorized researcher Maisie requests access to the email corpus.
10. Lynda loads the MBOX files (set 2) into DArcMail on a stand-alone workstation.
11. Lynda and Jennifer review the email accounts and messages and redact any sensitive or personal information that they find.
12. Maisie conducts her work and selects messages to be copied for further research.
13. Using DArcMail, Lynda generates a dissemination package (DIP) of Maisie's selected email messages by selecting and exporting the MBOX files (set 3). Lynda then copies them to a clean flash drive (provided by the Smithsonian Institution for security) for Maisie (figure 6).

5.3.3 Emulation Workflow Scenario

To illustrate the relevance and importance of software preservation and emulation in facilitating successful technical approaches to email archiving, we have outlined each step in retrieving a message from a user's inbox and the associated applications, then making email available in an emulated environment (figure 7).

Donor Relations and Pre-acquisition Appraisal

1. Jane, digital archivist at Institution X, is invited to accompany her director to an on-site consultation of a former information science researcher, Arkady Ivanov, who is thinking of donating his materials.
2. After discussing archival policies and processes around preservation and access, Mr. Ivanov is intrigued and shows Jane a project entitled VirtualMe from the late 2000s in which he purchased a laptop and proceeded to create a digital persona that existed entirely in virtual space as evidenced by digital traces that a living person would leave behind.
3. Mr. Ivanov agrees to donate his older computers and personal papers. While he wants many of the personal digital materials to remain restricted for a period of time, he is interested in

providing immediate access to all of the contents of VirtualMe, which includes a desktop Outlook instance.

Accession and Processing

4. After returning to the archives, Jane creates an accession record in ArchivesSpace and initiates a new resource description.
5. Jane takes the VirtualMe laptop to the BitCurator workstation and creates a disk image of the VirtualMe drive using Guymager, during which time she stores the accession number and resource description unique ID in the INFO file.
   a. Jane runs the BitCurator Reporting Tool and Disk Image Access Tool to investigate the contents.
   b. Jane uses the DFXML file output from BitCurator, which provides a listing of all files on the drive, including software executables and their dependencies.
   c. Jane prepares for emulated access to the email archive and associated attachments. She runs a script that compares the files on the drive and their checksums with those in a software metadata repository that stores descriptive and technical metadata about known or collected software and its checksums.
   d. Jane then compares the list of verified software (using checksums) against the list of file extensions and identifies any additional software that will be needed to render all files within the image.

Ingest and Access

6. Jane's institution uses an ingest-to-discovery workflow that stitches together several open-source tools.

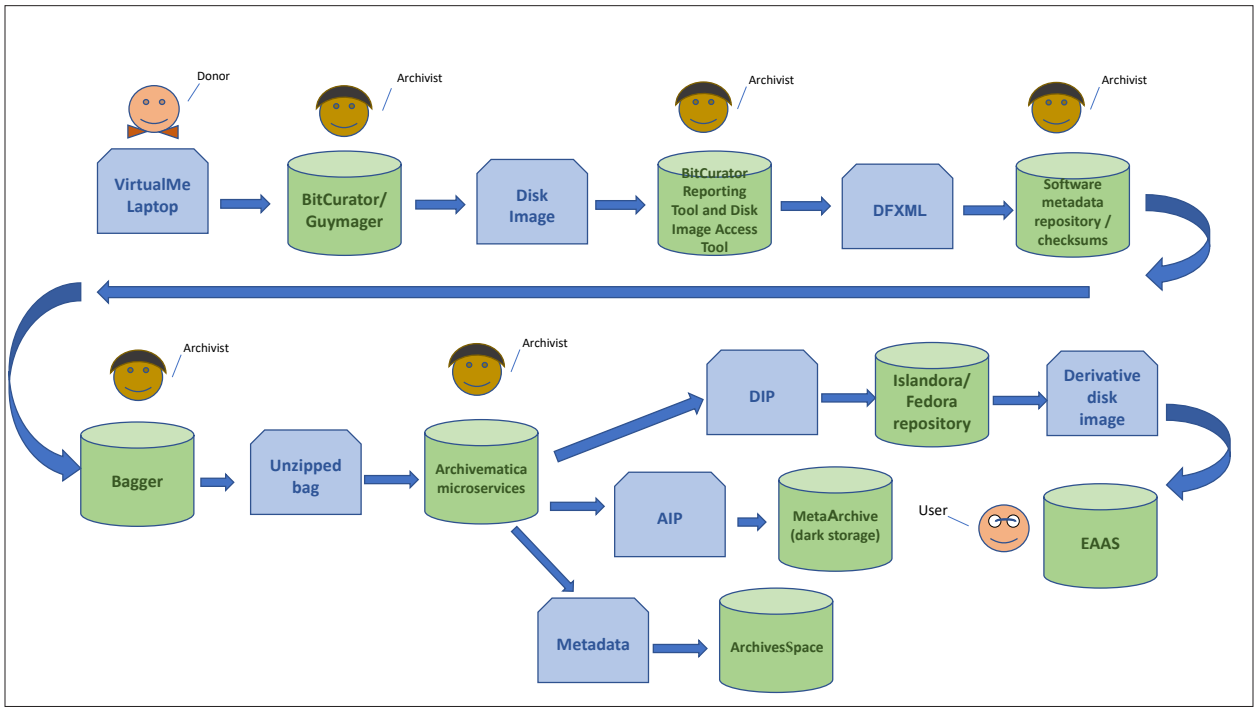

Fig. 7: Emulation workflow example: accessing a disk image of email and attachments

   a. Jane takes the output from BitCurator (including the disk image of the VirtualMe drive, BitCurator reports, bulk_extractor outputs, and DFXML file) and uses Bagger (GUI for the BagIt library) to create an unzipped bag that can be transferred through Archivematica microservices; metadata is added into a METS file at several points along the way.
   b. Archivematica generates a Dissemination Information Package (DIP), also an unzipped bag, which is routed to the Islandora/Fedora digital repository for management and discovery.
   c. Upon ingest into the appropriate collection, the VirtualMe bag is analyzed; a PID (process identification number) is assigned to the new object and several of the datastreams, including a METS file with accession number and resource description ID; and DFXML are indexed by Solr to allow for search and downstream machine actions.
   d. Archivematica generates an Archival Information Package (AIP), also an unzipped bag, which is routed to MetaArchive for dark archival preservation storage.
   e. Accession number and ArchivesSpace resource description IDs are used to pipe disk image data to ArchivesSpace so that a link to the object in Islandora and a minimal, descriptive entry for the disk image based on the METS file are added to the resource description inventory.
7. Remote user of the Islandora/Fedora repository
   a. Selects/clicks on the disk image in the Islandora discovery environment, which creates a derivative of the disk image with all additional software required to access all files (including email attachments) within the image.
   b. Uses a browser to access Emulation as a Service (EAAS), which makes the derivative image available.
   c. Views the entire contents and can produce DOIs for citations that represent the state of the machine at any point of access.

6. The Path Forward: Recommendations and Next Steps

Email represents an increasingly important part of the historical record. Preserving and ensuring access to this record is therefore central to the functions and values of archives and archivists (Society of American Archivists 2011). Until we can meet the challenges of email archiving, responsible custody is undermined, accountability is abandoned, and, ultimately, the historical record is imperiled. In short, the problem won't take care of itself, and the time to take action is now.

The workflow scenarios outlined in the previous section demonstrate that it is possible, but still difficult, for archival repositories to appraise, acquire, process, preserve, and provide access to email-based collections. Repository staff must choose from a range of tools, then connect them into often complicated workflows. While this is feasible for relatively well-resourced institutions with tech-savvy staff, most are being left behind. This is not because existing tools

cannot preserve email collections, but because the problem is complex. The community and tools are developing but are not yet fully mature. In some cases, basic research and policy decisions are still to be completed.

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 on 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, and projects that require more planning and funding.

6.1 Community Development and Advocacy

The most important work to be done in advancing email preservation lies simply in nurturing and fostering archives and libraries that are leading the work or wish to become more fully engaged. In theory, every archive that is collecting or preserving contemporary record material is collecting and preserving 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. A number of low-barrier activities could be pursued through existing groups. However, some external support and encouragement would also help the community coalesce around certain tools and services, providing the most sustainable long-term trajectory for essential email preservation technologies. Accordingly, the task force also recommends a few higher-impact activities that could be pursued in partnership with organizations supporting the cultural heritage community.

6.1.1 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 website (National Digital Stewardship Alliance 2013).
- **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.

Develop Training and Skills. 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, archivists and curators need to win the trust of donors and their organizations; this means showing specific ways that they can manage email in a responsible fashion. By the same token, organizational leaders must understand 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 the arrival of email collections.

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 History Lab or the University of Waterloo's Web Archives for Historical Research Group, may offer potential models for specialized training and information sharing. Such groups seem more likely to succeed once basic training is in place.

Activity: Develop a scalable training and workshop curriculum addressing the basics of email archives, including an overview of the issues and demonstration of available open-source (and potentially proprietary) tools. A half-day session will serve as a primer to email preservation; this report can be used 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 present an email archives tutorial at iPres 2018 in Boston and will release the training materials after the meeting; feedback will be incorporated into a revised version submitted for the International Council on Archives (ICA) 2019 Annual Conference in Edinburgh. Moving forward, regularized and more sustainable training could be developed for potential integration with existing curricula, such as those supported by the Society of American Archivists, Council of State Archivists' State Electronic Records Initiative, and the National Association of Government Archives and Records Administrators.

Demystify Email Archiving for Collection Donors. Donors of private digital collections are often confused about the importance of including email as a documentation source and about assurances of privacy and security.

Activity: Develop a customizable template for donor agreements that describes in detail the roles and responsibilities of both the donor and institutional repository, including information on

workflows, sensitivity review, redaction capabilities, potential embargo periods, and search and access.

Planned action: Members of the Email Task Force will put out a call for participation and convene 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: ePADD/Stanford will convene a working group from the user community to develop programming (which may include documentation or video tutorials) for account creators and donors.

Maintain Assessment of Email Tools in COPTR. The Email Task Force identified and analyzed common tools for email archiving. Because software tools are dynamic, with functionality added and subtracted regularly, this information should be stored in a flexible environment with wide public access.

Activity: Move tools list to the 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. If format migration is part of the preservation workflow, what are the advantages and consequences of 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 that compare and contrast format options for still images and reformatted video.[34] Once completed, the community would maintain the matrix through a public resource such as NDSA or the Digital Preservation Coalition.

Activity: Develop a format assessment matrix that includes information about the format's structure, standards documentation, technical metadata, header fields, expected behavior in common tools, and more.

Planned action: This work would incorporate results from the Test Existing Tools for Data Impact and Data Loss project and could be taken on as a follow-up by those involved in that project.

[34] FADGI Federal Agencies Digitization Guidelines Initiative, "Guidelines: File Format Comparison Projects–Still Image and Audio-Visual Working Groups," December 2, 2014, http://www.digitizationguidelines.gov/guidelines/File_format_compare.html.

6.1.2 High-Impact/Long-Term Activities

Sustain the Email-Archiving Community. The momentum generated by the task force's work makes this a good time to investigate steps 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 by consortia, but began their lives as research and grant-funded projects; well-known examples include BitCurator, CollectionsSpace, and ArchivesSpace.[35] While the software is freely available, an institution must join the consortium if it wants to support development or receive support.

The email-archiving community does not seem 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, and its open-source code is available in GitHub. TOMES, which is beginning to make its code available on GitHub, began its development a bit later and is more suitable for institutional archives (Gibson 2018). Both projects rely on grant funding for continued development. The Smithsonian Institution Archives makes code for DarcMail available on its website, and it will also be accessible on GitHub, as will that from EAS. During the task force discussions, members from these and other projects were keen to share information and experiences. This momentum should be encouraged as a complement to the specific tool development and implementation projects discussed in 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 needed for particular tools and services, as well as steps to build an organization dedicated to longer-term support.

Planned action: Develop complete project description, and seek funding for a project to define high-level functional needs for a more unified email capture, processing, and access tool.

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 capturing email from Capstone accounts or manager roles, or email related to particular case files.

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

Potential activity: The community might sponsor a summit or short project, perhaps in conjunction with the National Association of State Chief Information Officers (NASCIO), the Council of State Archivists (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 action: 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, NARA, and NASCIO. Such a project may benefit from external support to convene working meetings.

Develop Criteria for Email Authenticity. More exploration and documentation are 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 the development of a profile or schema to be used in validating the authenticity of specific properties of a message or account.

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 time? Is an email message 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? How can such factors be better captured to allow users to understand and then interpret the layers of evidence that may provide greater certainty that a message has been unaltered?

Activity: In 2012, the InSPECT Project developed a testing process to define the significant properties of email messages, then determine whether they were conserved when exported or migrated from a few test systems, including Thunderbird and Outlook (Knight 2010). The basic methodology was sound, but the work should be brought up-to-date to recognize new tools and evolution in email formats (such as 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 with practitioners.

Demonstrate Value for Email as Research Data Source. While there have been research projects that used publicly available email collections such as the Enron set, more work is needed to further the case that email is a rich source of research data for historians and others. Connecting digital humanities researchers, historians, and data scientists with full access to specific email collections will allow these researchers to conduct cross-disciplinary research on individuals and organizations in ways that are possible only through email collection analysis. Better understanding by the historical community could lead to JISC or other bodies studying the potential integration of email research into services such as JISC's Research Data Discovery.

Activity: Develop a "data challenge" project where historians and others apply to be "embedded scholars" at specific host institutions willing to give full access to email data sets for research. The researcher would publish, along with the results, specific reasons why the research findings would not be possible without access to email. Potential collaborators include the HILT Institute (Humanities Intensive Learning and Teaching) and American Historical Association.

Improve the IETF RFC Standards Documentation for MBOX. The current version of 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 necessarily compatible with those of 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 the process for revising a published RFC. Contact original RFC 4155 authors and other potential contributors to form a working group for revising the specification.

Improve Standards Documentation for EML. The EML format is only partially documented through IETF RFC 5322 for Internet Message Format (IMF). While IMF defines the ASCII text–based syntax for all email messages, the EML format is a subtype of IMF used by Microsoft Outlook Exchange and 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 a 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 are 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 simplify workflows at a large scale.

Activity: Work with the 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 the fall of 2018.

6.2 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 they depend largely on support from their parent institutions (and to a lesser extent, from partner repositories) or funding from federal granting agencies, whose forward funding is uncertain. In addition, repositories frequently rely on commercial tools 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 because of 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 should support small and large collections alike (both collections covering many accounts and those covering single accounts with many messages or attachments). In essence, archivists need email archiving tools with the ability to scale up or down as necessary, since what is large today will not be large tomorrow. The following recommendations are directed to the software development community as well as funders.

6.2.1 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 than 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, followed by 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 needs 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: 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 data sets 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. To promote true interoperability, it is important to integrate these capabilities within existing tools or within closely aligned applications rather than create stand-alone instances for each function. This would be a follow-on activity to the Test Existing Tools for Data Impact and Data Loss project (see previous recommendation).

Activity: After completing 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.

6.2.2 High-Impact/Long-Term Activities

Improve Tools for Sensitivity Review. One of the most pressing needs facing 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 such as Social Security numbers and phone numbers, it does not extend to less structured information such as education records (covered by FERPA[36]) or health records (covered by HIPAA[37]).

Natural language processing tools used for email review should be enhanced or complemented by machine-learning software to improve the ability of collections managers to identify and extract more nuanced entities from the archive. Current natural language processing workflows rely on named entity recognition to identify just certain data types, such as persons, corporations, and places, even offering some comparisons against specific categories in Wikipedia.

[36] Family Educational Rights and Privacy Act (FERPA): https://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html.

[37] Health Insurance Portability and Accountability Act of 1996 (HIPAA): https://www.hhs.gov/hipaa/for-professionals/security/laws-regulations/index.html.

(ePADD finds matches or "close" matches in email accounts. A next step might be integration with Wikidata categories and expansion of same.) The development of additional lexicons, such as those supported by ePADD and TOMES, will increase the effectiveness of human-in-the-loop processing. While machine learning is not foolproof, its use for facial recognition systems demonstrates its accuracy for large complex data sets, and there is no reason to think that it cannot be applied to email (Phillips 2018). And the presumption that human review of messages for sensitive content is better 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.

The much broader declassification and legal communities are allies in this need for better machine-learning options. 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" with regard 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 public interest activities" (Public Interest Declassification Board, 2012, 26). Similarly, Article 17 of the General Data Protection Regulation (GDPR) includes the Right to Erasure or Right to Be Forgotten. Its requirements allow data subjects to have the data controller erase their personal data, to cease further dissemination of the data, and potentially to have third parties halt processing of the data. The full effects of this regulation on email archives are not yet fully understood, but these provisions could impose significant requirements on archives for sensitivity review and redaction. 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 with studies done on technology-assisted review 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 could be identified and requirements established.

Activity: North Carolina State Archives is considering work on machine-learning tools (using Google's TensorFlow tools) to assist with classification and review in a 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.

Planned action: Several Email Task Force members are interested in developing a project plan and applying 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 natural language processing.

Sustain and Integrate 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 (and the almost infinite number of ways email can be rendered or displayed to users), it is not surprising that a wide range of workflows and tools are being employed. Yet, as noted in the discussion of workflows and implementation scenarios, 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 and 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 deserve 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, and others is needed to help close the gaps in current workflows and ensure better interoperability between tools. Could some of the tools be integrated into existing consortia or projects? If not, and the trend continues wherein different organizations manage different tools, how difficult will it become for each institution to administer these tools 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: At Stanford University, practitioners have mentioned the need for an aggregated discovery site, among other pressing needs for future development. This site 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 staff are investigating internal options to host such a site. Results from the Email Research Data Challenge project would help lay the groundwork for making email collections known, open, and available. 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.[38] TOMES is seeking to build connections to additional state and university archives, as well as to collecting repositories.

Activity: Since the summer of 2017, Harvard Library has been working on redeploying its email-processing tool EAS as open-source software. Progress to date includes the completion of a roadmap and technical plan, and preliminary technical work to enable the extraction of code that integrates closely with other institutionally supported systems. As currently resourced, it may be a couple of years before a version is available to the broader community. With additional IT funding, Harvard could increase the speed of development and hasten the deployment of a useful open-source tool. At the same time, Harvard would like to work with existing open-source projects to ensure useful compatibility between tools.

Activity: The Smithsonian Institution Archives has made an enhanced version of DArcMail available as of December 2017. The processing and preservation tool now supports SQLite as a database platform in addition to MySQL. This change improves its usefulness in organizations with limited staff and IT resources. The tool and its documentation are available as open source on the Smithsonian Institution Archives website.

Planned action: 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 the support of parent institutions, partner repositories, and external supporters.

Develop a 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 have a need or desire to keep a record of their own digital footprint, not unlike people of past generations, who stashed letters in a drawer, without taking the immediate step of donating their material 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 an unmet demand for a service that would allow for self-directed email capture for preservation, discovery, and

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

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 anyone with whom he or she wishes to share it. Such a service could even be designed to allow the eventual donation to a repository via an export feature or other method.

Activity and planned action: 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 numerous 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.

Standards succeed only 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 is 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 and 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 determine the core standards required to support interoperability between email-archiving tools and implement those standards in a select set of tools.

Appendix A:

Automating System Processes

Harvard Library automated the capture of system processes to record provenance, maintaining records that map to the Event entity in the PREMIS data model. Metadata is recorded at the item/message level and at the packet level. "Packet" is the unit of ingest into Harvard's Email Archiving System (EAS) and roughly corresponds to a collection, as shown in the following graphic.

Events in EAS

Associated docs	g:\LTS\LDI Systems\EAS\specs\deletions_wg_20120416_PM.docx

Events in EAS are a means of tracking the processing history of items in EAS.

Current events created in EAS are:

Event Level	Event Content Type	Event Type	Event Value	Agent
Item (1)	email message	normalization	extracted from a file in {0} format and converted it to {1} format	Emailchemy
			e.g. *extracted from a file in Outlook for Windows/unknown format and converted it to RFC-2822 eml format*	
Item	email message	normalization	rewrote line breaks	EAS
Item (2)	email message	normalization	removed embedded files	EAS
Item	email message	association	associated external attachment [{0}]	EAS
Item	email message	delete component	deleted attachment [{0}]	person
Item (3)	email message	metadata update	AdminFlag removed: [{0}]	person
Item	email message attachment	normalization	extracted from a file in {0} format and converted it to {1} format	Emailchemy
			e.g. *extracted from a file in Outlook for Windows/unknown format and decoded*	
Item	email message attachment	delete component	deleted email message [{0}]	person
Packet	email message	delete component	deleted email message [{0}]	person
Packet	email message attachment	delete component	deleted attachment [{0}]	person

Notes:

1. This event modifies the email message by adding a header e.g.:

```
X-Converted-By: Emailchemy 11.2.2 Embedded Edition; licensedTo="Harvard_Library_1"
```

2. This event modifies the email message by adding a header:

```
X-Converted-By: EAS 0.9.2; conversionDate=2013-01-09T19:20:07 UTC
```

It also removes the embedded attachment and modifies the embedded content header and content e.g.:

```
--------------070006080000030409010704
Content-Type: text/plain; charset=UTF-8
X-EAS-INLINE-ID: 970492
X-EAS-MESSAGE-DIGEST: type=md5;0b330856810e35dd98289225bceffed1
```

```
Start EAS original body part headers
Content-Type: image/png; name="pre_processing_flow.png"
Content-Transfer-Encoding: base64
Content-Disposition: inline; filename="pre_processing_flow.png"
End EAS original body part headers
```

```
EAS Inline Converted: extracted to external location; pre_processing_flow.png; cid:
```

3. This event is not written to DRS as an event but is written to `hulDrsAdmin/adminFlag/flagID`

Appendix B:

Email Tools for Libraries, Archives, and Museums

While the private sector and records management needs have long been significant drivers in the development of email tools, the past decade has witnessed a concerted effort by libraries, archives, and museums (LAM) to implement workflows and develop new resources that specifically address key portions of the email stewardship lifecycle model. Their efforts have been intended to ensure that electronic correspondence of administrative or historical value is preserved and remains accessible over the long term. In addition to repurposing and incorporating email tools in digital preservation workflows, cultural heritage institutions have developed new tools tailored to their specific functional requirements and at the same time have embarked on numerous projects to refine methodologies and establish best practices for email preservation. Developers in the LAM community have also availed themselves of email-specific libraries in programming languages (such as Python's Email and Mailbox modules) that can be used to parse, manipulate, and transform messages and folders of correspondence to satisfy preservation requirements. Recent years have also witnessed the emergence of digital preservation systems such as Preservica and Archivematica, which provide comprehensive workflows that take content from the point of acquisition through deposit in a preservation repository. They employ a microservice design so that specific functions (and associated tools) can be introduced into workflows to address key needs. While proprietary tools (and systems, such as Preservica) are used across the LAM sector, the aforementioned advances in email preservation among institutions are highlighted by an embrace of open-source technology that can be modified and adapted to fulfill unique requirements and integrate component tools into larger workflows and systems.

A more complete list of email tools, including tools from outside the cultural heritage domain, is available at http://www.emailarchivestaskforce.org/documents/email-tools/.

Archivematica

Basics

Developed by: Artefactual Systems

Links:

- https://www.archivematica.org/en/
- https://wiki.archivematica.org/Main_Page

Availability: System is freely available and open source. All Archivematica code is released under a GNU Affero General Public License (AGPL 3.0).

System requirements:

- OS:
  - Ubuntu 14.04.5 64 bit Server Edition
  - CentOS 7.3.1611 64 bit
  - Support for Ubuntu 16.04 is planned for a future release.
- Hardware requirements (minimum production requirements):
  - Processor: 2 CPU cores
  - Memory: 4 GB+
  - Disk space: 20 GB plus three to four times the disk space required for the collection being processed (e.g., 200 GB to process a 50 GB transfer)

Status: Actively developed and maintained by Artefactual Systems, with additional contributions from the growing user community. Version 1.7 was released in May 2018.

What does the tool do?

Purpose: Archivematica employs a microservice design to "provide an integrated suite of software tools that allows users to process digital objects from ingest to access in compliance with the ISO-OAIS functional model." Furthermore, it employs METS and PREMIS to record and track descriptive, administrative, and rights metadata. In the initial transfer stage, content is placed into the Archivematica backlog after being passed through a workflow that includes the generation of file UUIDs and checksums; quarantine and virus scans; file format identification, characterization, and validation; technical metadata extraction; and deployment of a forensics tool (bulk_extractor) to identify sensitive or private information. The appraisal and arrangement stage includes functionality to help archivists review and analyze the technical aspects and informational content of digital archives. Integration with ArchivesSpace furthermore permits the creation of archival description and its association with digital content. The ingest stage prepares Archival Information Packages and Dissemination Information Packages for deposit to appropriate platforms and includes a workflow step for the normalization of files for preservation and access.

Position in stewardship lifecycle: Processing and preservation. In terms of accessioning, Archivematica permits normalization from native email formats (for a small number of source formats) as well as bulk processing to accomplish such tasks as file format identification, characterization, and validation; technical metadata extraction; and identification of personal or sensitive information. Archivists can perform intellectual arrangement (for instance, associating MBOX or EML files with elements of archival description) and also package content for deposit to a preservation repository (with any such storage locations tracked via Archivematica's associated Storage Service application). Archivematica itself is not involved with online discovery or access.

Strengths, weaknesses, and gaps: Among its strengths are Archivematica's microservice design and pipeline architecture, which allow archivists to establish set workflows that move content through a standardized set of procedures. The system includes numerous essential preservation actions (with an exhaustive audit trail of preservation metadata) that will help ensure content retains its authenticity and integrity over time. At the same time, the system has some gaps in terms of handling email, which include a failure to separate and separately preserve message attachments and normalization pathways for a very small subset of email formats. In this regard, it might be appropriate to employ Archivematica in tandem with another tool (such as ePADD), which could perform more advanced, email-centric operations and would then feed content into the transfer and ingest pipelines to complete standardized preservation actions. Archivematica likewise lacks more advanced search and data analysis functionality.

Formats

Ingest:

- MBOX
- PST
- Maildir

Other email formats may be ingested by Archivematica but will receive only default preservation actions.

Export:

- MBOX
- Original file formats

DArcMail (Digital Archive Mail System)

Basics

Developed by: Smithsonian Institution Archives (SIA); maintained by the archives and user community

Link: http://siarchives.si.edu/blog/yes-we%E2%80%99re-still-talking-about-email

Availability: A user guide and download are available from links at https://siarchives.si.edu/what-we-do/digital-curation/email-preservation-cerp.

System requirements:

- OS: Windows, Mac, or Linux
- Python
- Relational databases: MySQL or SQLite
- Available in workstation/client or server versions

Status: Currently in version 1. The Smithsonian currently maintains and supports it, but is looking forward to the time when it is not the only code contributor.

What does the tool do?

Purpose: The DArcMail tool is designed to be used for initial appraisal and then for preservation (Archival Information Package [AIP]) and access (Dissemination Information Package [DIP]). It natively retains the logical arrangement of the original account in both the AIP and DIP packages. Its flexibility allows for the creation of custom subsets of email for creation of specialized AIPs and DIPs. (Simpson 2016).

Position in stewardship lifecycle: Processing, preservation, and discovery and access. DArcMail provides normalization, item level and bulk processing, intellectual arrangement, search capability, packaging, and access functionality for email.

Strengths, weaknesses, and gaps: Users may process emails at varying levels (individual messages, a group thereof, or entire accounts), and the tool may be used to process up to 100,000 in a batch. The platform converts MBOX files to EMail Account XML (EMA), a comprehensive schema that fulfills RFC 5322 preservation requirements and may be applied to everything from a single message to an entire account. Each message and account embedded in the EMA preservation format receives an SHA-1 checksum. At the same time, the tool does not normalize to EML, and the EMA schema has not been widely adopted. Furthermore, additional metadata (such as rights) must be created and recorded in a separate system.

Formats

Ingest: MBOX

Export:

- XML email schema (AIP master copy)
- MBOX (DIP access copy)
- Attachments embedded or separated into parallel structure

EAS (Electronic Archiving System)

Basics

Developed by: Harvard Library at Harvard University

Link: http://nrs.harvard.edu/urn-3:hul.eother:eas_overview

Availability: The system is highly integrated with Harvard's technical infrastructure and is currently available to the Harvard community only; work to make EAS open source began in the summer of 2017.

System requirements: EASi, the staff administrative user interface for EAS, is a web-based application currently available to authorized users in the Harvard community (via ID and PIN). Because EAS manages potentially sensitive data, it requires a designated secure VPN tunnel. Supported browsers are Firefox, Google Chrome, and Safari.

Status: The system is actively maintained and supported. Support is currently provided to the Harvard community only. Version 1.0.5 was released in April 2017.

What does the tool do?

Purpose: EAS is a processing tool for archivists that supports ingest of email content (email and attachments) for appraisal, processing, and deposit to Harvard's preservation repository. EAS supports working with email content in batches or individually: adding, editing, and removing metadata; deleting individual email messages or attachments; marking content for future review; marking content that requires secure storage for sensitive data; associating PREMIS rights metadata; and creating an association with a collection. An additional feature is that EAS automatically records technical metadata and system events such as format conversion, deletions, and the handling of attachments. All metadata is stored at the item level (an email message or attachment).

Position in stewardship lifecycle: Processing and preservation. EAS is a processing tool for ingest, archival appraisal and processing, and packaging for deposit to a preservation repository.

Strengths, weaknesses, and gaps: The strength of EAS is in its flexibility to accommodate multiple levels of processing; entire collections can be deposited directly to the preservation repository for long-term storage with minimal metadata, or metadata can be applied to individual email messages and attachments. The robust search interface allows archivists to refine search criteria and to see how many instances of a particular metadata value exist within a selected set. EAS has no provision for pre-acquisition appraisal for archivists. Once the content and metadata have been deposited to Harvard's preservation repository, only authorized account users can access it. Because of embargo periods and concerns about sensitive data, at this time there is no public discovery or delivery to end users. To enable pre-acquisition appraisal, discovery, and delivery functions, EAS needs to be used in a workflow with other tools (e.g., ePADD). To make it possible for those at Harvard and in the broader community to use EAS with other tools, a project to deploy EAS as open-source software began in the summer of 2017.

Formats

Ingest: OLM, MBOX, PST, and EMLX. The tool is currently configured to import email content (messages and attachments) from specific clients, with new ones added as needed. EAS currently supports email content from the following:

- Eudora for Windows/6.2
- Eudora for Windows/version unknown
- Mac OS X Mail/2.x
- Mailman/2.0.5

- Mailman/2.1.15
  - Outlook for Mac (OLM only/version unknown)
  - Outlook for Windows/version unknown
  - Thunderbird/2.0.0.23
  - Thunderbird/version unknown

All email is converted to EML for processing and preservation storage. The line breaks of the EML are normalized by EAS. Attachments are decoded and stored as separate files within EAS, with metadata relationships to associated email. For email originating from certain email clients that store email separately from attachments, EAS will try to match the email with its associated attachments via metadata relationships.

Export: There is no export from EAS (except within Harvard's infrastructure to the preservation repository).

ePADD (Email: Process, Appraise, Discover, Deliver)

Basics

Developed by: Stanford University Libraries, with technical support from Sudheendra Hangal, faculty member in computer science at Ashoka University

Links:

- http://epadd.stanford.edu/epadd/collections
- https://github.com/ePADD/epadd/releases

Availability: Open-source and licensed under an Apache public license, v2.0; current version is available at GitHub—ePADD 5.1 at https://github.com/ePADD/epadd/releases/.

System Requirements:

- OS: 64-bit, Windows 7 SP1 / 10, Mac OS X 10.11 / 10.12
  - Windows installations: Java Runtime Environment 64-bit, 8u101 or later required
  - Optimized for Windows 7 and OSX 10.9/10.10 machines, using Java 7 or 8
- Browser-based software client is compatible with Chrome and Firefox.
- Memory: 8 GB RAM (4 GB allocated to the application by default)
- ePADD is written in Java and Javascript and powered by Apache Tomcat (v7.0) using Java EE Servlet API (v3.x) and Java Mail (v1.4.2). Text and metadata extraction, indexing, and retrieval are performed by Apache Lucene (v4.7) and Apache Tika (v1.8). Charting and visualization are supported using the D3-based reusable chart library (v0.4.10). Oracle's Java Application Bundler and Launch4J are used for packaging on Mac and Windows platforms, respectively. Other Java libraries from Apache (e.g., Lang, commons, CLI, IO, logging) are also used. JSON formatting is performed with the libraries org.json and Gson.

- ePADD has implemented its own natural language processing (NLP) toolkit, which is used for named entity extraction, disambiguation, and other tasks. This toolkit supplants the Apache OpenNLP used in earlier beta versions of the ePADD software. Stanford continues to use Muse as an internal library within ePADD. However, the Apache OpenNLP proved insufficient for their needs (at least for name recognition), and after various rounds of customization, they built a bespoke named-entity recognizer. This toolkit uses external data sets such as Wikipedia/DBpedia, Freebase, Geonames, OCLC FAST and LC Subject Headings/LC Name Authority File.

Status:

- Currently in active development, ePADD is managed by Stanford University's Department of Special Collections & University Archives, in collaboration with partners at Harvard University, the Metropolitan New York Library Council (METRO), University of Illinois at Urbana-Champaign, and University of California, Irvine.
- Funding for current ePADD development is provided through an Institute of Museum and Library Services (IMLS) National Leadership Grant for Libraries, which supports projects that address challenges faced by the library and archival fields and that have the potential to advance practice in those fields. Development for the initial 2015 release of ePADD was primarily funded by the National Historical Publications and Records Commission (NHPRC).
- User documentation maintained by Stanford University Libraries is available at https://docs.google.com/document/d/1joUmI8yZEOnFzuWaVN1A5gAEA8UawC-UnKycdcuG5Xc/edit.
- Information on active user community forums, the mailing list, focus group meetings, and the lexicon working group is available at https://library.stanford.edu/projects/epadd/community.

What does the tool do?

Purpose: ePADD is a software package developed by Stanford University Special Collections & University Archives. This tool supports appraisal by donor or curator, processing, discovery (online publication of metadata), and delivery of full-text unrestricted messages and attachments in a reading room environment.

Position in stewardship lifecycle: ePADD includes four modules—Appraisal, Processing, Discovery, and Delivery—which are designed to facilitate the process of working with email archives at the following stages of the lifecycle:

- Appraisal and selection: The Appraisal module allows creators, dealers, and curators to easily gather and review email archives prior to transferring those files to an archival repository.
- Archival processing: The Processing module provides archivists with the means to arrange and describe email archives.

  - Discovery and access for research: The Discovery module provides the tools for repositories to remotely share a redacted view of email archives with users through a web server discovery environment. The Delivery module enables archival repositories to provide moderated full-text access to unrestricted email archives within a reading room environment.

Strengths, weaknesses, and gaps: ePADD is still in development with additional releases scheduled by the end of 2018.

- Strengths:
  - Turns unstructured data to structured data automatically
  - Links extracted entities to permanent identifiers automatically
  - Provides relevant authority records for users to confirm the relationship
  - Allows annotations to email messages
  - Allows browsing of all attached images in one place
  - Groups different email addresses belonging to same persons
  - Groups entities
  - Generates queries for users
  - Provides templates to create complex search as lexicon
  - Allows separate email messages from mailing list
- Gaps:
  - In the Discovery environment, cross collection searching should be enhanced.
  - In the Delivery environment, not all file formats of attachments are viewable—only those that Quick View Plus can render. In addition, users should be allowed to annotate or correct metadata, such as correspondent names, etc.
  - Exports of GraphML files are not yet supported for social networking.
  - At this point, the attachments are not transformed and would need to be coupled with a commercial software (like Quick View Plus) with the Delivery Module to view/render attachments in over 300 obsolete file formats.

Formats

Ingest: The import screen opens in the Appraisal module (default). This interface is where information (name and an associated email address) is entered about the owner of the email account. If applicable, it is also where the location and account information can be specified for the MBOX files or IMAP email accounts that ePADD will be ingesting for review and potential transfer to an archival repository. Multiple accounts, as well as multiple MBOX files, can be selected.

Export: ePADD currently exports MBOX for preservation repositories. Attachments can be selected for export according to various search criteria from the exports screen, with options to export only those attachments that have not been recognized by Apache Tika (and are therefore not indexed with ePADD), for further review; parameters that use specifies in the Discovery module (version 3.0); metadata extracted via NLP for publication in ePADD's online Discovery environment; and CSV files for authorities.

Preservica Standard Edition

Basics

Developed by: Preservica, Inc. (a subsidiary company backed by Tessella Archiving Solutions)

Links:

- http://preservica.com/
- Email Archiving and Preservation webinar recording (requires registration with Citrix GoToWebinar) https://preservica.com/events/webinars-live-demos/15/07/2015/email-archiving-and-preservation

What does the tool do?

Purpose: The product website previously noted that "Preservica provides a comprehensive suite of OAIS (Open Archival Information System) compliant workflows for ingest, data management, storage, access, administration and preservation, as well as our new Universal Access module that allows [institutions] to safely share open content [...] with the public." Preservica provides fully automated ingest procedures (with the capability to bulk upload collections of 10+ TB); advanced users can employ a Submission Information Package creator tool. The microservice design of Preservica pushes content through a standardized workflow that includes steps such as virus scans, checksum calculation, file characterization, and technical metadata extraction in addition to file format normalization. Archivists may add metadata to digital content, which can be searched via the curatorial interface and Universal Access module.

Position in stewardship lifecycle:

- Archival processing: While it does not support pre-acquisition appraisal, Preservica normalizes email to the EML format and facilitates archival processing, with item- (message and attachments) and batch-level processing, the maintenance of "conversational relationships" to assist with intellectual arrangement, and indexing of message contents.
- Preservation: Preservica also assists with packaging content and storing it in a repository.
- Discovery and access for research: Preservica assists with online discovery and access via its Universal Access module.

Strengths, weaknesses, and gaps: While not an email-specific tool, Preservica has notable features that would contribute to the long-term preservation and dissemination of any such content. The automated workflow that encompasses key preservation actions (including format normalization) is a key feature, and the ability to extract metadata (including specific information about the relationship of email messages to larger threads or folders) or add description are important for archival processes. The integrated nature of the platform is also significant, in particular the fact that it can support both

the preservation of and access to email content. Preservica does not appear to include sensitive/personal information identification or more advanced search capabilities (such as natural language processing or named entity recognition).

Formats

Ingest:

- MS Outlook (PST and MSG)
- Lotus Notes
- MBox
- Gmail (via export to MBOX with Google Takeout)

Export:

- EML
- Attachments may be normalized to preservation formats.
- Metadata can be exported to EAD, MODS, or Dublin Core, as well as to systems such as Axiell Calm, Adlib, and ArchivesSpace.

TOMES Tool (Transforming Online Mail with Embedded Semantics)

Basics

Developed by: State Archives of North Carolina, with support from the State Archives of Utah and the Kansas State Historical Society

Links:

- https://www.ncdcr.gov/resources/records-management/tomes
- https://github.com/StateArchivesOfNorthCarolina

Availability: Open-source, current version available at GitHub.

System requirements:

- Any 64-bit OS that can run Docker; see https://www.docker.com/community-edition
- 4 GB RAM minimum
- A modern browser (Chrome, Firefox, IE, Safari)

Status: Currently in active development, TOMES is managed by the State Archives of North Carolina. Funding for current TOMES development is provided through an NHPRC State Electronic Records grant through September 2018.

What does the tool do?

Purpose: The TOMES tool allows archivists to process complete email accounts more quickly by using NLP tagging to identify personally identifiable information, confidential information, and named entities. It uses dictionaries specific to state government in an XML format. TOMES is still in development, with release planned for late September 2018.

Position in stewardship lifecycle: Acquisition and archival processing

Strengths, weaknesses, and gaps:

Strengths:

- Assists in processing very large email accounts typically found in government records contexts
- Allows iterative processing, so that difficult-to-process accounts may be made accessible faster
- Depends only on Docker so it can be implemented in many computing environments

Gaps:

- TOMES NLP assistance features can require specialized knowledge to build into effective libraries.
- EAXS is a limited framework for preservation and should be reconceptualized for a modern email processing.

Formats

Ingest: PST, MBOX, EML

Export: EAXS XML with tagging

Appendix C:
Email Preservation Research Projects

Archiving Email Symposium

Investigators: Library of Congress and the National Archives and Records Administration

Overview: "On June 2, 2015, the Library of Congress and the National Archives and Records Administration co-hosted the Archiving Email Symposium at the Library to share information about the state of practice in accessioning and preserving email messages and related attachments. The approximately 150-person audience included a wide range of practitioners, from technologists and software developers, librarians, curators, records managers, lone arranger archivists and academics, and representatives from large federal agencies with many thousands of employees as well as grant funding programs, including the National Endowment for Humanities, Institute of Museum and Library Services and National Historical Publications and Records Commission. In addition, the event included an informal workshop on June 3 with a subset of participants to discuss issues and challenges identified during the Symposium in order to better define the gaps in our tools, processes and policies for archiving email collections." (Murray and Engle 2015)

Date: 2015

Reports and Resources:
- http://www.digitalpreservation.gov/meetings/archivingemailsymposium.html
- We Welcome Our Email Overlords: Highlights from the Archiving Email Symposium

Carcanet Press Email Preservation Project

Investigator: University of Manchester Library

Overview: "Among the most important modern archives held by the University of Manchester Library (UML) is that of Carcanet Press, one of the UK's premier poetry publishing houses. Correspondence with famous poets, critics, editors, translators and artists forms one of the most important elements of this archive. Most of this correspondence is now conducted by email, with the result that the quantity of hard copy correspondence acquired in annual accruals to the archive has diminished significantly. It is therefore vital that libraries such as the UML are able to preserve these emails in digital form.

This will ensure that invaluable primary research material is not lost to the archival record" (Baker, Butler, and Green 2012, 3). The Carcanet Press Email Preservation Project used "both traditional archival practice and digital preservation standards" to produce a number of significant outcomes, which included

- software code for metadata extraction and for automatic verification of migration experiments
- a full metadata profile and a data model for Archival Information Packages
- new curatorial documentation
- dedicated digital preservation hardware and a secure network drive for initial processing of digital archives

Date: 2012

Reports and Resources:

- https://www.escholar.manchester.ac.uk/api/datastream?publicationPid=uk-ac-man-scw:165096&datastreamId=FULL-TEXT.PDF
- https://www.escholar.manchester.ac.uk/api/datastream?publicationPid=uk-ac-man-scw:226625&datastreamId=FULL-TEXT.PDF

CERP (Collaborative Electronic Records Project)

Investigators: Smithsonian Institute Archives and the Rockefeller Archive Center

Overview: The Collaborative Electronic Records Project (CERP) initially sought "to develop the methodology and technology for managing and preserving born-digital materials in archival collections. The project's primary objectives were to produce management guidelines and technical preservation capability that would enable archives and manuscript repositories to make electronic information accessible and usable for future researchers, and to share findings and products with depositors, peer institutions, and other interested non-profit groups." The project's scope soon narrowed to focus on email, given its ubiquity and associated preservation challenges. The project team collaborated with the Electronic Mail Collection and Preservation (EMCAP) initiative to develop an email account XML schema and by its 2008 conclusion, "CERP had produced best practices guidelines, a workflow outline, evaluation of software tested, SIP/AIP/DIP models, a software tool that preserves email accounts together with their messages (the CERP parser), and a customized DSpace ingest module, and had parsed more than 89,000 email messages with a success rate of 99 percent" (Adgent and Fuhrig 2009, 3–4).

Dates: 2005–2008

Reports and Resources:
- http://siarchives.si.edu/cerp/
- http://siarchives.si.edu/cerp/
- CERP_project_summary_122008_CC.pdf

DAVID (Digital Archiving in Flemish Institutions and Administrations)

Investigator: Filip Boudrez (Stadsarchief Antwerpen)

Overview: "DAVID, Digital Archiving in Flemish Institutions and Administrations, is a project of the Foundation for Scientific Research within the scope of the Max Wildiers Foundation and is a cooperation between Antwerp City Archives and the Interdisciplinary Centre for Law and Informatics of the K.U. Leuven. The goal of this project was to create a manual on electronic archiving. . . . The DAVID project examined the judicial and archival requirements for e-mail preservation and pointed out some possible archiving strategies. On this basis, a model solution was developed. In addition to the theoretical concept, this report also contained an initial incentive for the practical implementation of a records management and record-keeping procedure for e-mails and related electronic documents." (Boudrez 2006, 2)

Dates: 1999–2003

Reports and Resources:
http://www.imaginar.org/taller/dppd/DPPD/179%20pp%20DAVID.pdf (version 2)

Kaine Email Project@LVA

Investigator: Library of Virginia

Overview: In January 2010, the administration of outgoing Governor Tim Kaine transferred to the Library of Virginia approximately 1.3 million email messages from more than 200 email accounts. By law, gubernatorial records transferred to the Library "shall be made accessible to the public, once cataloging has been completed" (Va. Code § 2.2-126). The Kaine Email Project required the Library of Virginia to establish new tools and procedures to accession, process, and provide access to the governor's electronic correspondence.

Dates: 2010–Ongoing

Reports and Resources:
http://www.virginiamemory.com/collections/kaine/

MeMail (Email Preservation at the University of Michigan)

Investigator: University of Michigan Bentley Historical Library

Overview: A two-year project funded by The Andrew W. Mellon Foundation, the Bentley Historical Library's MeMail Project sought to overcome the preservation challenges posed by diverse email applications and personal email management practices at the University of Michigan by including record creators in the appraisal and selection of email of long-term value and identifying appropriate tools to facilitate the transfer of electronic correspondence to the archives. Pilot participants would drag/drop, forward, or copy messages of value to "archival mailboxes" established by archivists, who would then export the email and move it through a preservation workflow. The project ultimately found that record creators were unable to appraise and select email of value with confidence and the archival mailbox transfer method proved unsustainable when the university adopted Gmail as an email service. At the same time, the project did help the Bentley establish more robust digital preservation procedures and workflows.

Dates: 2010–2011

Reports and Resources:

- SAA Campus Case Study #14: Partnering with IT to Identify a Commercial Tool for Capturing Archival Email of University Executives at the University of Michigan http://files.archivists.org/pubs/CampusCaseStudies/CASE-14-FINAL.pdf
- SAA Campus Case Study #15: Will They Populate the Boxes? Piloting a Low-Tech Method for Capturing Executive E-mail and a Workflow for Preserving It at the University of Michigan http://files.archivists.org/pubs/CampusCaseStudies/CASE-15-FINAL.pdf

PeDALS (Persistent Digital Archives and Library System)

Investigators:

- Arizona State Library Archives and Public Records (lead institution)
- Alabama Department of Archives and History
- State Library and Archives of Florida
- New Mexico State Records Center and Archives
- New York State Archives, New York State Library
- South Carolina Department of Archives and History and South Carolina State Library
- Wisconsin Historical Society

Overview: "The Persistent Digital Archives and Library System, or PeDALS, was a research project, from January 2008 to March 2012, which had two technical goals. First, was to develop a curatorial rationale to support an automated, integrated workflow to process collections of digital publications and records. Second, to implement 'digital stacks' using an inexpensive, storage network that can preserve the authenticity and integrity of the collections. In addition to those technical goals, PeDALS sought to build a community of shared practice so that the system meets the needs of a wide range of repositories that could then support the ongoing development of the system and promote best practices. To further that end, PeDALS strove to remove barriers to adopting the technology by keeping costs as low as possible." (PeDALS 2013)

Dates: 2008–2012

Reports and Resources:
- http://web.archive.org/web/20130306060835/http://www.pedalspreservation.org/
- http://azmemory.azlibrary.gov/cdm/ref/collection/statepubs/id/15540

TOMES (Transforming Online Mail with Embedded Semantics)

Investigators: State Archives of North Carolina, Utah Division of Archives and Records Service, Kansas Historical Society

Overview: "The Transforming Online Mail with Embedded Semantics (TOMES) project, generously funded by the National Historical Publications and Records Commission, seeks to identify email accounts of public officials with enduring value in order to capture, preserve and provide access to important government records. TOMES is a multi-state partnership that includes Kansas, Utah and North Carolina focused on developing processes for transferring email accounts out of hosted email solution platforms, e.g. Microsoft 365 and Gmail, and converting them into a sustainable open source language. Additionally, the team will build on the work of e-PADD to develop an appraisal tool using natural language processing and a state government specific dictionary to aid archivists to quickly process and provide access." (NCDNCR 2018)

Dates: 2015–2018

Reports and Resources:
https://www.ncdcr.gov/resources/records-management/tomes

Appendix D:
Reference List

All URLs were current as of May 31, 2018.

Adgent, Nancy, and Lynda Schmitz Fuhrig. 2009. *The Collaborative Electronic Records Project Summary*. Sleepy Hollow, NY, and Washington, DC: The Collaborative Electronic Records Project. http://siarchives.si.edu/cerp/CERP_project_summary_122008_CC.pdf.

AIIM. 2018. Glossary. Accessed May 31. http://www.aiim.org/Resources/Glossary/Glossary-List-Page.

Alderman, Liz. 2017. "Bell Pottinger, British P.R. Firm for Questionable Clients, Collapses." *New York Times*, September 12 (Business Day). https://www.nytimes.com/2017/09/12/business/bell-pottinger-administration.html.

Animal Adventure Park. 2018. Official Animal Adventure Park April the Giraffe Page. Accessed May 31. http://www.aprilthegiraffe.com/.

Ankerson, Megan. 2012. "Writing Web Histories with an Eye on the Analog Past." *New Media and Society* 14 (3): 384–400. https://doi.org/10.1177/1461444811414834.

Apple. 2018. "iCloud: Add an Email Attachment in iCloud Mail." Posted March 30, 2018. https://support.apple.com/kb/ph2629?locale=en_US.

Archives New Zealand. 2018. "Visual Rendering Matters." Accessed May 31. http://archives.govt.nz/resources/information-management-research/rendering-matters-report-results-research-digital-object-0.

Attfield, Simon, and Larry Chapin. 2018. "The Reconstruction of Narrative in E-Discovery Investigations." Presentation at Email Preservation: How Hard Can it Be? January 24, 2018, Woburn House, London. Digital Preservation Coalition. https://www.dpconline.org/docs/miscellaneous/events/2018-events/1766-dpc-email-ii-attfield-chapin/file.

AV Preserve. 2018. "Fixity." Accessed May 31. https://www.avpreserve.com/products/fixity/.

Baker, Fran. 2014. *Carcanet Press Email Preservation Project Phases 2-3: Final Report*. Carcanet Press Email Preservation Project. United Kingdom: Carcanet Press. https://www.escholar.manchester.ac.uk/api/datastream?publicationPid=uk-ac-man-scw:226625&datastreamId=FULL-TEXT.PDF.

____ . 2015. "E-Mails to an Editor: Safeguarding the Literary Correspondence of the Twenty-First Century at The University of Manchester Library." *Special Collections in a Digital Age* 21 (2): 216–224. https://doi.org/10.1080/13614533.2015.1040925.

Baker, Fran, Phil Butler, and Ben Green. 2012. *Carcanet Press Email Preservation Project: JISC Final Report*. https://www.escholar.manchester.ac.uk/api/datastream?publicationPid=uk-ac-man-scw:165096&datastreamId=FULL-TEXT.PDF.

Banday, M. T. 2011. "Technology Corner: Analysing E-Mail Headers for Forensic Investigation." *Journal of Digital Forensics, Security and Law* 6 (2): 49–64. https://doi.org/10.15394/jdfsl.2011.1095.

Bearman, David. 2017. "Review of 'Office of the Secretary: Evaluation of Email Records Management and Cybersecurity Requirements, ESP-16-03.'" *The American Archivist* 80 (2): 459–462. https://doi.org/10.17723/0360-9081-80.2.459.

Ben-David, Anat, and Hugo Huurdeman. 2014. "Web Archive Search as Research: Methodological and Theoretical Implications." *Alexandria: The Journal of National and International Library and Information Issues* 25 (1/2): 93–111. https://doi.org/10.7227/ALX.0022.

BITS Security Program. 2013. *Email Authentication Policy and Deployment Strategy for Financial Services Firms*. Washington, DC: BITS Security Working Group. http://www.fsroundtable.org/wp-content/uploads/2015/05/BITSEmailAuthenticationFeb2013.pdf.

Boudrez, Filip. 2006. *Filing and Archiving Email*. http://www.expertisecentrumdavid.be/docs/filingArchiving_email.pdf.

Bradner, S. 1996. "The Internet Standards Process—Revision 3. RFC–Best Current Practice." Harvard University. https://datatracker.ietf.org/doc/rfc2026/.

Brunton, Finn. 2013. *Spam: A Shadow History of the Internet*. Cambridge, MA, and London: The MIT Press. https://mitpress.mit.edu/books/spam.

BSG Web Group. 2017. "Wrench." September 5, 2017. https://nasa3d.arc.nasa.gov/detail/wrench-mis.

Bunn, Jenny, Sara Brimble, Selene Obolensky, and Nicola Wood. 2015. *Team Europe EU28 Project 2015–16: Perceptions of Born Digital Authenticity*. InterPARES Report. InterPARES Trust. https://interparestrust.org/assets/public/dissemination/EU28_20160718_UserPerceptionsOfAuthenticity_FinalReport.pdf.

Caplan, Priscilla. 2009. *Understanding PREMIS*. Library of Congress Network Development and MARC Standards Office. Washington, DC: Library of Congress. https://www.loc.gov/standards/premis/understanding-premis.pdf.

Casserly, Martyn. 2017. "The Best Free Email Services for 2017." *Tech Advisor*. Internet Feature (blog), September 11, 2017. https://www.techadvisor.co.uk/feature/internet/best-free-email-services-for-2017-3613837/.

Cerbain, Jose. 2016. "Reports of the Death of Email Are Greatly Exaggerated." *Huffington Post* (blog), July 2, 2016. https://www.huffingtonpost.com/advertising-week/reports-of-the-death-of-e_b_11114786.html.

Cheney, Kyle. 2017. "GOP Chairmen Seek to Interview Top FBI Officials on Clinton, Trump." *Politico*, December 19. https://www.politico.com/news/hillary-clinton-emails.

Cocciolo, Anthony. 2016. "Email as Cultural Heritage Resource: Appraisal Solutions from an Art Museum Context." *Records Management Journal* 26 (1): 68–82. http://dx.doi.org/10.1108/RMJ-04-2015-0014.

Cohen, Jordan. 2015. "Looking at the EMAIL MARKETINGscape" (blog), January 26, 2015. https://www.emailvendorselection.com/author/jordan-cohen/.

Cohen, William W. 2015. Enron Email Dataset. May 8, 2015. https://www.cs.cmu.edu/~./enron.

Context.io. 2018. Accessed May 31. https://docs.context.io/.

Cormack, Gordon V., and Maura R. Grossman. 2017. "Navigating Imprecision in Relevance Assessments on the Road to Total Recall: Roger and Me." In *SIGIR '17 Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 5–14. ACM Press. https://doi.org/10.1145/3077136.3080812.

Crispin, Mark R. 2003. "Internet Message Access Protocol–Version 4 Rev1. RFC 3501." Network Working Group, Internet Engineering Task Force. https://tools.ietf.org/html/rfc3501.

Crocker, D. 2008. "Internet Mail Architecture." Network Working Group, Internet Engineering Task Force. http://www.bbiw.net/specifications/draft-crocker-email-arch-11.html.

Daintith, John, and Edmund Wright. 2008. *A Dictionary of Computing*. 6th ed. New York: Oxford University Press, Inc.

Dappert, Angela, Sébastien Peyrard, Carol C. H. Chou, and Janet Delve. 2013. "Describing and Preserving Digital Object Environments." *New Review of Information Networking* 18 (2): 106–173. https://doi.org/10.1080/13614576.2013.842494.

Dayley, Alan, Julian Tirsu, Garth Landers, and Shane Harris. 2016. *Magic Quadrant for Enterprise Information Archiving*. ID: G00294240. Stamford, CT: Gartner, Inc. https://www.gartner.com/doc/3535317.

Digital Curation Centre. 2018. "DCC Curation Lifecycle Model." Accessed May 31. http://www.dcc.ac.uk/resources/curation-lifecycle-model.

Digital Preservation Coalition. 2015. "File Formats and Standards." *Digital Preservation Handbook*, 2nd ed. https://www.dpconline.org/handbook/technical-solutions-and-tools/file-formats-and-standards.

dmarc.org. 2018. Domain Message Authentication Reporting & Conformance (DMARC). Accessed May 31. https://dmarc.org/.

Ducheneaut, Nicolas, and Victoria Bellotti. 2001. "E-Mail as Habitat: An Exploration of Embedded Personal Information Management." *Interactions* 8 (5) 30–38. https://dl.acm.org/citation.cfm?id=383305.

Duke Law Center for Judicial Studies. 2018. "New EDRM Enron Email Data Set." Accessed May 31. https://www.edrm.net/resources/data-sets/edrm-enron-email-data-set/.

Dyer, Jessica. 2017a. "UNM's Krebs Encouraged Staffers to Purge Emails." *Albuquerque Journal*, October 29. https://www.abqjournal.com/1085003/unms-krebs-encouraged-staffers-to-purge-emails.html.

____ . 2017b. "Krebs Said He Deleted UNM Emails." *Albuquerque Journal*, November 23. https://www.abqjournal.com/1097004/krebs-said-he-deleted-unm-emails.html.

Email Design Reference. 2018. "Responsive Email." MailChimp. Accessed May 31. https://templates.mailchimp.com/development/responsive-email/.

Exterro. 2018. "Predictive Coding (Technology Assisted Review)". *The Basics of E-Discovery*, Chap. 7B. Accessed May 31. https://www.exterro.com/basics-of-e-discovery/predictive-coding/.

Ferguson, Rob. 2017. "Civil Servants Kept Emails in Case They Shed Light on McGuinty's Gas Plant Cancellations." *The Toronto Star*, October 20 (Queen's Park). https://www.thestar.com/news/queenspark/2017/10/20/emails-kept-in-case-they-shed-light-on-mcguintys-gas-plant-cancellations.html.

Ferrante, Ricc. 2015. "Archiving Email: Institutional Approaches to Processing and Archiving Email [902]." Presentation at Archiving Email Symposium, June 2, 2015, Washington, DC, co-hosted by the Library of Congress and the National Archives and Records Administration. https://www.youtube.com/watch?v=4xTVnkqsOF0&list=PLEA69BE43AA9F7E68&t=0s&index=5. Transcript available at https://stream-media.loc.gov/webcasts/captions/2015/150602osi0902.txt.

Firstbrook, Peter, and Neil Wynne. 2015. *Magic Quadrant for Secure Email Gateways*. ID: G00268427. Stamford, CT: Gartner, Inc. https://www.gartner.com/doc/3084025/magic-quadrant-secure-email-gateways.

Fitzgerald, Neal. 2013. "Using Data Archiving Tools to Preserve Archival Records in Business Systems–A Case Study." http://purl.pt/24107/1/iPres2013_PDF/Using%20data%20archiving%20tools%20to%20preserve%20archival%20records%20in%20business%20systems%20%E2%80%93%20a%20case%20study.pdf.

Freed, N., and N. Borenstein. 1996. "Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies. RFC 2045." Network Working Group, Internet Engineering Task Force. https://tools.ietf.org/rfc/rfc2045.

FWD:Everyone. 2018. Accessed May 31. https://www.fwdeveryone.com/.

Gearan, Anne, and Philip Rucker. 2017. "Trump Criticizes How Mueller Obtained Transition Emails, Says No Plans to Fire Special Counsel." *Washington Post*, December 17 (Politics). https://www.washingtonpost.com/politics/mueller-unlawfully-obtained-emails-trump-transition-team-says/2017/12/16/6162f350-e2cc-11e7-8679-a9728984779c_story.html.

Gibson, Jeremy. 2018. "State Archives of North Carolina." GitHub. Accessed May 31. https://github.com/StateArchivesOfNorthCarolina.

Gmail Help. 2018. "Send Google Drive Attachments in Gmail-Computer." Gmail Help. Accessed May 31. https://support.google.com/mail/answer/2487407?co=GENIE.Platform%3DDesktop&hl=en.

Google. 2017. "Gmail API Overview." Last updated May 11, 2017. https://developers.google.com/gmail/api/guides/.

Grace, Stephan, Gareth Knight, and Lynn Montague. 2009. *Investigating the Significant Properties of Electronic Content Over Time. Final Report.* https://web.archive.org/web/20151024064901if_/http://www.significantproperties.org.uk/inspect-finalreport.pdf.

Graham, Robert. 2016. "Yes, We Can Validate the Wikileaks Emails." *Errata Security* (blog), October 21, 2016. http://blog.erratasec.com/2016/10/yes-we-can-validate-wikileaks-emails.html.

Grammer, Geoff. 2017. "Krebs' Emails Show His Desire to Protect Donors, Acknowledges Mistakes Made." *Albuquerque Journal*, September 18. https://www.abqjournal.com/1065745/krebs-emails-show-his-desire-to-protect-donors-acknowledges-mistakes-made.html.

Grossman, Maura R., and Gordon V. Cormack. 2011. "Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient than Exhaustive Manual Review." *Richmond Journal of Law and Technology* 17 (3): 1–48. http://jolt.richmond.edu/jolt-archive/v17i3/article11.pdf.

____ . 2014. "Comments on 'The Implications of Rule 26(g) on the Use of Technology-Assisted Review.'" *The Federal Courts Law Review* 7 (1): 285–313. http://www.fclr.org/fclr/articles/pdf/comments-implications-rule26g-tar-62314.pdf.

Harvard Library. 2016. "Email Archiving Stewardship Workshop." Published March 16, 2016. http://library.harvard.edu/03092016-1642/email-archiving-stewardship-workshop.

Harvard University Archives. 2018. "University Records Policies." Accessed May 31. http://library.harvard.edu/university-archives/managing-university-records/policies.

Harvard Wiki. 2018. *DRS Content Guide to In-Production Content Models*. Last modified January 26, 2018. https://wiki.harvard.edu/confluence/pages/viewpage.action?pageId=204385879&preview=/204385879/218248076/public_drs_content_guide.pdf.

History Lab. 2018. "History as Data Science (Freedom of Information Archive)." Accessed May 31. http://www.history-lab.org.

hMailServer. 2018. Storing the Message in the Database. Accessed January 10. https://www.hmailserver.com/documentation/latest/?page=faq_storing_message_in_database.

Hopkins, Julie, and Adam Sarner. 2015. *Market Guide for Email Marketing*. ID: G00276045. Stamford, CT: Gartner, Inc. https://www.gartner.com/doc/3173144?ref=SiteSearch&sthkw=email%20marketing&fnl=search&srcId=1-3478922254.

Huth, Geof. 2016. "Appraising Digital Records." In *Appraisal and Acquisition Strategies*, edited by Michael Shallcross and Christopher Prom, 7–68. Chicago: Society of American Archivists.

Illinois State Archives, and Records and Information Management Services, University of Illinois at Urbana-Champaign. 2017. "Processing Capstone Email Using Predictive Coding." University of Illinois, Urbana Champaign. https://www.uillinois.edu/cio/services/rims/about_rims/projects/processing_capstone_email_using_predictive_coding/.

Informatica. 2018. "Email Verification." Accessed May 31. https://www.informatica.com/products/data-quality/data-as-a-service/email-verification.html#fbid=q2fSFYMpuEz.

Jääskeläinen, Anssi, Miia Kosonen, and Lissa Uosukainen. 2017. "Developing a Citizen Archive." BloggERS! (blog), February 1, 2017. https://saaers.wordpress.com/2017/02/01/developing-a-citizen-archive/.

John, Jeremy Leighton, Ian Rowlands, Peter Williams, and Katrina Dean. 2010. *Digital Lives: Personal Digital Archives for the 21st Century: An Initial Synthesis*. A Digital Lives Research Paper, beta version 0.2 (March 3). British Library. http://britishlibrary.typepad.co.uk/files/digital-lives-synthesis02-1.pdf.

Johnston, David. 1990. "5,000 Files Erased From Poindexter's Computer." *New York Times*, March 16. http://www.nytimes.com/1990/03/16/us/5000-files-erased-from-poindexter-s-computer.html.

Kekre, Anand. 2015. "Securing Email Attachments with Digital Rights Management." *Vaultize's Secure Enterprise File Sharing Blog*, April 16, 2015. http://www.vaultize.com/blog/securing-email-attachments-with-digital-rights-management.

Klensin, J. 2008. "Simple Mail Transfer Protocol. RFC 5321." Network Working Group, Internet Engineering Task Force. http://tools.ietf.org/html/rfc5321.

Knight, Gareth. 2010. *Significant Properties Testing Report: Electronic Mail*. Jisc, The National Archives, and Kings College London. https://web.archive.org/web/20151024134638if_/http://www.significantproperties.org.uk/email-testingreport.pdf.

Kucherawy, M. 2009. "Message Header Field for Indicating Message Authentication Status. RFC 5451." Network Working Group, Internet Engineering Task Force. https://tools.ietf.org/html/rfc5451.

Landers, Garth, Shane Harris, and Jie Zhang. 2017. *When to Use Microsoft's Native Capabilities for Archiving and E-Discovery*. ID: G00315437. Stamford, CT: Gartner, Inc. https://www.gartner.com/doc/3658817.

Larramo, Mika. 2018. "SMTP Commands Reference." Accessed May 31. http://www.samlogic.net/articles/smtp-commands-reference.htm.

Lavoie, Brian, and Richard Gartner. 2013. *Preservation Metadata,* 2nd ed. DPC Technology Watch Report 13. Great Britain: Digital Preservation Coalition. http://www.dpconline.org/docs/technology-watch-reports/894-dpctw13-03/file.

Leiba, Barry. 2013. "Change the Status of ADSP (RFC 5617) to Historic." Internet Engineering Task Force. https://datatracker.ietf.org/doc/status-change-adsp-rfc5617-to-historic/.

Levinson, Edward, ed. 1998. "Content-ID and Message-ID Uniform Resource Locators. RFC 2392." Network Working Group, Internet Engineering Task Force. https://tools.ietf.org/html/rfc2392.

Library of Congress. 2016. "MBOX Email Format." Last significant FDD update November 17, 2016. https://www.loc.gov/preservation/digital/formats/fdd/fdd000383.shtml.

Library of Congress. 2018. "Digital Content Transfer Tools." Digital Preservation. Accessed May 30. http://www.digitalpreservation.gov/series/challenge/data-transfer-tools.html.

Library of Virginia. 2016. "Kaine Email Project @ LVA." Virginia Memory. http://www.virginiamemory.com/collections/kaine/.

Light, Michelle, and Tom Hyry. 2002. "Colophons and Annotations: New Directions for the Finding Aid." *The American Archivist* 65 (2): 216–230. https://doi.org/10.17723/aarc.65.2.l3h27j5x8716586q.

Lin, Tom C. W. 2016. "Compliance, Technology, and Modern Finance." *Temple University Legal Studies Research Paper No. 2017-06*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2904664.

Loftus, Mary J. 2010. "The Author's Desktop." *Emory Magazine* (Winter). http://www.emory.edu/EMORY_MAGAZINE/2010/winter/authors.html.

MacAskill, Ewen, and Owen Bowcott. 2017. "UK Prosecutors Admit Destroying Key Emails in Julian Assange Case." *The Guardian*, November 10 (Media). http://www.theguardian.com/media/2017/nov/10/uk-prosecutors-admit-destroying-key-emails-from-julian-assange-case.

Maemura, Emily, Christoph Becker, and Ian Milligan. 2016. "Understanding Computational Web Archives Research Methods Using Research Objects." *Proceedings of IEEE International Conference on Big Data*, December 5–8, 2016, Washington, DC. https://doi.org/10.1109/BigData.2016.7840982.

Maildir. 2018. *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Maildir&oldid=836696378.

Manjoo, Farhad. 2017. "What We Lose When the World Moves On From Email." *New York Times*, July 12. https://www.nytimes.com/2017/07/12/technology/what-we-lose-when-the-world-moves-on-from-email.html?_r=1.

Marshall, Catherine C. 2008a. "Rethinking Personal Digital Archiving, Part 1: Four Challenges from the Field." *D-Lib Magazine* 14 (3/4). http://www.dlib.org/dlib/march08/marshall/03marshall-pt1.html.

____ . 2008b. "Rethinking Personal Digital Archiving Part 2: Implications for Services, Applications, and Institutions." *D-Lib Magazine* 14 (3/4). http://www.dlib.org/dlib/march08/marshall/03marshall-pt2.html.

Microsoft. 2005. "Exchange Storage Architecture." Microsoft Exchange Server 2003. https://technet.microsoft.com/en-us/library/bb124808(v=exchg.65).aspx.

____ . 2011. "Compliance Features in Exchange Online." Last modified December 19, 2011. https://msdn.microsoft.com/en-us/library/hh147162(v=exchsrvcs.149).aspx.

Microsoft Developer Network. 2011. "Set Up and Manage Information Rights Management in Exchange Online." Microsoft Exchange Server. https://msdn.microsoft.com/en-us/library/gg597271(v=exchsrvcs.149).aspx.

Microsoft Developer Network. 2018. "Outlook Mail REST API Reference." Last updated April 5, 2018. https://msdn.microsoft.com/en-us/office/office365/api/mail-rest-operations.

Microsoft Exchange Online. 2016. "Journaling in Exchange Online." https://technet.microsoft.com/en-us/library/jj898487(v=exchg.150).aspx.

____ . 2017. "In-Place Hold and Litigation Hold." https://technet.microsoft.com/en-us/library/ff637980(v=exchg.150).aspx.

Mimecast. 2018. "E-Discovery & Compliance." Accessed January 11. https://www.mimecast.com/solutions/archive/e-discovery-and-compliance/.

Morgan, Steve. 2015. "Cybersecurity Market Reaches $75 Billion In 2015; Expected To Reach $170 Billion By 2020." *Forbes*, December 20. https://www.forbes.com/sites/stevemorgan/2015/12/20/cybersecurity%E2%80%8B-%E2%80%8Bmarket-reaches-75-billion-in-2015%E2%80%8B%E2%80%8B-%E2%80%8Bexpected-to-reach-170-billion-by-2020/#1d2d65a930d6.

Murray, Kate. 2014. "Shaking the Email Format Family Tree." *The Signal* (blog), April 4, 2014. Library of Congress. http://blogs.loc.gov/thesignal/2014/04/shaking-the-email-format-family-tree/.

Murray, Kate, and Erin Engle. 2015. "We Welcome Our Email Overlords: Highlights from the Archiving Email Symposium." *The Signal* (blog), July 9, 2015. Library of Congress. https://blogs.loc.gov/thesignal/2015/07/we-welcome-our-email-overlords-highlights-from-the-archiving-email-symposium/?loclr=blogsig.

NARA (National Archives and Records Administration). 1997. "Resources–Publications: Disposition of Federal Records–Chapter 6." In *Disposition of Federal Records: A Records Management Handbook*, 2nd ed.

Washington, DC: Office of Records Services. https://www.archives.gov/records-mgmt/publications/disposition-of-federal-records/chapter-6.html#VI.PermanentRecords.

_____ . 2010. "Processing the Presidential Records of Elena Kagan." *AOTUS Blog: The Blog of the Archivist of the United States*, June 22, 2010. https://aotus.blogs.archives.gov/2010/06/22/processing-the-presidential-records-of-elena-kagan/.

_____ . 2013a. "Guidance on a New Approach to Managing Email Records." *NARA Bulletin* 2013-02. https://www.archives.gov/records-mgmt/bulletins/2013/2013-02.html.

_____ . 2013b. *User Guide: Managing NARA Email Records with Gmail and the ZL Unified Archive*, version 1.0. https://www.archives.gov/files/records-mgmt/email-management/sample-agency-user-guide-for-managing-email.pdf.

_____ . 2015. *White Paper on The Capstone Approach and Capstone GRS*. https://www.archives.gov/files/records-mgmt/email-management/final-capstone-white-paper.pdf.

_____ . 2018. *National Archives 2018–2022 Strategic Plan*. https://www.archives.gov/about/plans-reports/strategic-plan.

National Digital Stewardship Alliance. 2013. "Levels of Digital Preservation," version 1. http://ndsa.org//activities/levels-of-digital-preservation/.

NCDNCR (North Carolina Department of Natural and Cultural Resources). 2018. "Transforming Online Mail with Embedded Semantics (TOMES)." Accessed May 31. https://www.ncdcr.gov/resources/records-management/tomes.

Never Bounce. 2018. Accessed May 31. https://neverbounce.com.

NHPRC (National Historical Publications and Records Commission). 2018. "Transforming Online Mail with Embedded Semantics (TOMES)." Accessed May 31. https://www.ncdcr.gov/resources/records-management/tomes.

Novell Documentation. 2018. "Information Stored in the Post Office." *GroupWise 2014 R2 Administration Guide*. Accessed May 31. https://www.novell.com/documentation/groupwise2014r2/gw2014_guide_admin/data/adm_poa_understand_post_office_info.html.

Nuix. 2018. "EDRM Enron Data Set." Accessed May 31. https://www.nuix.com/edrm-enron-data-set/edrm-enron-data-set.

Nylas. 2018. "The Nylas APIs." Accessed May 31. https://docs.nylas.com/reference.

OAIS. 2012. "Space Data and Information Transfer Systems—Open Archival Information System (OAIS)–Reference Model." https://www.iso.org/obp/ui/#iso:std:iso:14721:ed-2:v1:en.

Oard, Douglas, William Webber, David Kirsch, and Sergey Golitsynskiy. 2015. "Avocado Research Email Collection—Linguistic Data Consortium." Release date: February 16, 2015. https://catalog.ldc.upenn.edu/ldc2015t03.

Oberhaus, Daniel. 2016. "How to Use the Internet on the Summit of Everest." *Motherboard* (blog), July 31, 2016. https://motherboard.vice.com/en_us/article/4xa4zp/when-the-internet-came-to-everest.

Onishi, Norimitsu. 2017. "South African Court Raises Pressure for Zuma to Go." *New York Times*, December 29 (Africa). https://www.nytimes.com/2017/12/29/world/africa/south-africa-court-zuma-impeach.html.

Organized Crime and Corruption Reporting Project. 2017. "#GuptaLeaks to Be Released to Journalists Worldwide." Press release, November 10. https://www.occrp.org/en/40-press-releases/pressss-releases/7240-guptaleaks-to-be-released-to-journalists-worldwide.

Owens, Trevor. 2014. "The EPADD Team on Processing and Accessing Email Archives." *The Signal* (blog), October 20, 2014. Library of Congress. http://blogs.loc.gov/thesignal/2014/10/the-epadd-team-on-processing-and-accessing-email-archives/.

Pace, Nicholas, and Laura Zakaras. 2012. *Where the Money Goes: Understanding Litigant Expenditures for Producing Electronic Discovery*. Santa Monica, CA: The RAND Corporation. http://www.rand.org/content/dam/rand/pubs/monographs/2012/RAND_MG1208.pdf.

Palarchio, Joe. 2015. "Office 365—Is the 'Archive Mailbox' Still Relevant?" *Perficient* (blog), June 29, 2015. https://blogs.perficient.com/microsoft/2015/06/office-365-is-the-archive-mailbox-still-relevant/.

Paradigm Project. 2008. *Workbook on Digital Private Papers*. https://www.webarchive.org.uk/wayback/archive/20080701041124/http://www.paradigm.ac.uk/workbook/index.html.

Pearce-Moses, Richard. 2005. *A Glossary of Archival and Records Terminology*. Chicago: Society of American Archivisits. http://files.archivists.org/pubs/free/SAA-Glossary-2005.pdf. See also https://www2.archivists.org/glossary.

PeDALS. 2013. "About PeDALS." Accessed via Wayback Machine March 6, 2013. http://web.archive.org/web/20130306060835/http://www.pedalspreservation.org/.

Pennock, Maureen. 2006. "Curating Emails: A Lifecyle Approach to the Management and Preservation of Email Messages." In *DCC Digital Curation Manual*, edited by Seamus Ross and Michael Day. http://www.dcc.ac.uk/sites/default/files/documents/resource/curation-manual/chapters/curating-e-mails/curating-e-mails.pdf.

Phillips, Tom. 2018 "China Testing Facial-Recognition Surveillance System in Xinjiang." *The Guardian*, January 18. https://www.theguardian.com/world/2018/jan/18/china-testing-facial-recognition-surveillance-system-in-xinjiang-report.

Pinpoint Labs. 2018. Accessed May 31. http://www.pinpointlabs.com/productsservices/software/harvester.

Plante, Jeanette. 2015. "Challenges of Email as a Record Archiving Email Symposium." Presentation at Archiving Email Symposium, Department of Justice, June 2, 2015. http://www.digitalpreservation.gov/meetings/documents/aes15/9_Plante_2015-06-02%20Library%20of%20congress%20email%20management.pdf.

Pogue, David. 2004. "State of the Art; Google Mail: Virtue Lies In the In-Box." *New York Times*, May 13. http://www.nytimes.com/2004/05/13/technology/state-of-the-art-google-mail-virtue-lies-in-the-in-box.html.

Pontevolpe, Gioanfranco, and Silvia Salsa. 2009. *General Study 05—Keeping and Preserving Email*. InterPARES 3 Project. http://www.interpares.org/ip3/display_file.cfm?doc=ip3_italy_gs05a_final_report.pdf.

Pratt, Kathryn Mary "Kary." n.d. "The Developing Standards for Authenticating Electronic Evidence." Accessed May 31. http://lawsdocbox.com/Legal_Issues/70550752-The-developing-standards-for-authenticating-electronic-evidence-kathryn-mary-kary-pratt.html.

PREMIS Editorial Committee. 2008. *PREMIS Data Dictionary for Preservation of Metadata*, version 2.0. http://www.loc.gov/standards/premis/v2/premis-2-0.pdf.

Princeton University Department of Rare Books and Special Collections. 2018. "Access Policy for University Archives Collections." Accessed May 31. https://rbsc.princeton.edu/policies/access-policy-university-archives-collections.

Prom, Christopher J. 2011. *Preserving Email*. Great Britain: Digital Preservation Coalition. http://dx.doi.org/10.7207/twr11-01.

Public Interest Declassification Board. 2012. *Transforming the Security Classification System*. Report to the President from the PIDB. https://www.archives.gov/files/declassification/pidb/recommendations/transforming-classification.pdf.

Radicati Group, Inc. 2016. "Email Market, 2016–2020—Executive Summary." London, UK: The Radicati Group, Inc. http://www.radicati.com/wp/wp-content/uploads/2016/01/Email_Market_2016-2020_Executive%20Summary.pdf.

Reagan Library. 2017. "Reagan Library Topic Guide—Iran-Contra Scandal." Reagan Library. https://www.reaganlibrary.gov/sites/default/files/archives/textual/topics/iran-contra.pdf.

Resnick, Peter W., ed. 2008. "Internet Message Format. RFC 5322." Network Working Group, Internet Engineering Task Force. https://tools.ietf.org/html/rfc5322.

Return Path. 2015. *Deliverability Benchmark Report: Analysis of Inbox Placement Rates in 2015*. https://returnpath.com/wp-content/uploads/2015/10/2015-Deliverability-Benchmark-Report.pdf.

Rockefeller Archive Center. 2006. *E-Mail Guidelines for Managers and Employees*. The Collaborative Electronic Records Project. http://www.nypap.org/wp-content/uploads/2016/04/rockefeller._email_guidelines.pdf.

Rockmore, Dan. 2014. "The Digital Life of Salman Rushdie." *The New Yorker*, July 29. https://www.newyorker.com/tech/elements/digital-life-salman-rushdie.

Rodden, Kerry, and Michael Leggett. 2010. "Best of Both Worlds: Improving Gmail Labels with the Affordances of Folders." In *CHI EA '10 CHI '10 Extended Abstracts on Human Factors in Computing Systems*, 4587–4596. Atlanta: ACM. https://doi.org/10.1145/1753846.1754199.

Rosica, Jim. 2017. "Despite 'Questions,' Grand Jury Clears Andrew Gillum in Email Controversy." *Florida Politics* (blog), August 8, 2017. http://floridapolitics.com/archives/242684-grand-jury-andrew-gillum.

Rothenberg, Jeff. 1999. *Ensuring the Longevity of Digital Information*. https://www.clir.org/wp-content/uploads/sites/6/ensuring.pdf.

Rothenberg, Jeff. 2000."Preserving Authentic Digital Information." In *Authenticity in a Digital Environment*, 51–68. Washington, DC: Council on Library and Information Resources. https://www.clir.org/pubs/reports/pub92/rothenberg/.

Rouse, Margaret. 2018. Definition: Electronic Discovery (e-discovery or ediscovery). Accessed May 31. http://searchfinancialsecurity.techtarget.com/definition/electronic-discovery.

Sabato, Larry. 1998. "The Iran-Contra Affair—1986–1987." Washingtonpost.com Special Report: Clinton Accused. http://www.washingtonpost.com/wp-srv/politics/special/clinton/frenzy/iran.htm.

Simpson, Joel. 2016. *Email Archiving Systems Interoperability*. Harvard Library Report. https://dash.harvard.edu/bitstream/handle/1/28682572/HL_Email_Archiving_Systems_Interoperability_Report_2016.pdf?sequence=3.

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

Spangler, Todd. 2017. "April the Giraffe's Live Birth Drew 14 Million YouTube Views on One Day." *Variety*, April 17. http://variety.com/2017/digital/news/april-giraffe-live-birth-youtube-1202032345/.

SparkPost. 2015. *Market Guide for Email Marketing*. Stamford, CT: Gartner, Inc. https://pages.messagesystems.com/Gartner-WP-Download-Landing-Page.html?src=Blog&pc=BL-GD-GartnerEmailGuide.

Stanford University. 2018. "EPADD-Documentation." Stanford Libraries. Accessed May 31. https://library.stanford.edu/projects/epadd/documentation.

Sustainability of Digital Formats. 2013a. "Microsoft Outlook PST 97-2002 (ANSI)." Library of Congress Digital Preservation. Last

significant FDD update November 25, 2013. http://www.digitalpreservation.gov/formats/fdd/fdd000377.shtml.

_____ . 2013b. "Microsoft Outlook PST 2003 (Unicode)." Library of Congress Digital Preservation. Last significant FDD update November 25, 2013. http://www.digitalpreservation.gov/formats/fdd/fdd000378.shtml.

_____ . 2015. "Lotus Notes Storage Facility." Library of Congress Digital Preservation. Last significant FDD update April 4, 2018. https://www.loc.gov/preservation/digital/formats/fdd/fdd000433.shtml.

Task Force on Technical Approaches for Email Archives. 2018a. "Exploring Email Emulation." http://www.emailarchivestaskforce.org/documents/exploring-email-emulation/.

_____ . 2018b. "Guide to Email Standards." http://www.emailarchivestaskforce.org/documents/guide-to-email-standards/.

_____ . 2018c. "Managing Email for Preservation." http://www.emailarchivestaskforce.org/documents/managing-email-for-preservation/.

_____ . 2018d. "Email User Features." http://www.emailarchivestaskforce.org/documents/email-user-features/.

_____ . 2018e. "Email Archiving Tools." http://www.emailarchivestaskforce.org/documents/email-tools/.

The Coalition of Technology Resources for Lawyers. 2016. *2016 Guidelines Regarding the Use of Technology-Assisted Review*. http://ctrlinitiative.com/wp-content/uploads/2014/07/2016-Guidelines-Regarding-the-Use-of-Technology-Assisted-Review.pdf.

The Document Foundation Wiki. 2018. "Feature Comparison: Mozilla Thunderbird–Microsoft Outlook." Last edited May 9, 2018. https://wiki.documentfoundation.org/Feature_Comparison:_Mozilla_Thunderbird_-_Microsoft_Outlook.

The National Archives. 2016. *The Application of Technology-Assisted Review to Born-Digital Records Transfer, Inquiries and Beyond*. United Kingdom: The National Archives. http://www.nationalarchives.gov.uk/documents/technology-assisted-review-to-born-digital-records-transfer.pdf.

Underwood, William, Marlit Hayslett, Sheila Isbell, Sandra Laib, Scott Sherrill, and Matthew Underwood. 2009. *Advanced Decision Support for Archival Processing of Presidential Electronic Records: Final Scientific and Technical Report*. Technical Report ITTL/CSITD 09-05. Atlanta, GA: Georgia Tech Research Institute. http://perpos.gtri.gatech.edu/publications/TR%2009-05-Final%20Report.pdf.

U.S. Government Accountability Office. 2008. "Federal Records: Agencies Face Challenges in Managing E-Mail," GAO-08-699T (April 23). https://www.gao.gov/products/GAO-08-699T.

U.S. Securities and Exchange Commission. 2002. "The Sarbanes-Oxley Act of 2002." https://www.sec.gov/about/laws/soa2002.pdf.

Vaudreuil, Gregory M., and Glenn Parsons. 2004. "Voice Profile for Internet Mail Version 2 (VPIMv2). RFC 3801." Network Working Group, Internet Engineering Task Force. https://tools.ietf.org/html/rfc3801.

Vogel, Karl, and Charles Cazabon. n.d. Database Back-End for Large Email Systems. Memoryhole. Accessed May 31, 2018. http://www.memoryhole.net/~kyle/databaseemail.html.

Waters, Donald, and John Garrett. 1996. *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information*. Washington, DC and Mountainview, CA: The Commission on Preservation and Access and The Research Libraries Group. https://www.clir.org/pubs/reports/pub63/.

Weiss, Debra Cassens. 2017. "Law Firm's Automatic Deletion of Spam Emails Is Blamed for Failure to File Timely Appeal." *ABA Journal*, September 28. http://www.abajournal.com/news/article/law_firms_automatic_deletion_of_spam_emails_is_blamed_for_failure_to_file_t/.

Whitt, Richard S. 2017. "'Through a Class Darkly' Technical, Policy, and Financial Actions to Avert the Coming Digital Dark Ages." *Santa Clara High Technology Law Journal* 33 (2): 117–229. http://digitalcommons.law.scu.edu/chtlj/vol33/iss2/1.

Wikipedia. 2016. *Lorraine v. Markel American Insurance Co.* https://en.wikipedia.org/w/index.php?title=Lorraine_v._Markel_American_Insurance_Co.&oldid=709391592.

_____ . 2017a. "Emulator." https://en.wikipedia.org/w/index.php?title=Emulator&oldid=815244333.

_____ . 2017b. "Comparison of Webmail Providers." https://en.wikipedia.org/w/index.php?title=Comparison_of_webmail_providers&oldid=816032285.

Zdziarski, Jonathan. 2008. "Email Database." In *iPhone Forensics*, 1st ed., 79. Sebastopol, CA: O'Reilly Media, Inc. https://books.google.com/books?id=R1XArTHPn9QC&pg=PA79&lpg=PA79&dq=how+are+email+messages+stored+in+database+tables?&source=bl&ots=_gzF-o6Fpsn&sig=NuJRndLZwmi7wAIvpuHTBd4oNbA&hl=en&sa=X&ved=0ahUKEwiH6-Dnrb3UAhUGPT4KHei8A4I4ChDoAQghMAA#v=onepage&q=how%20are%20email%20messages%20stored%20in%20database%20tables%3F&f=false.

Zhang, Jie, Debra Logan, and Garth Landers. 2014. *Magic Quadrant for E-Discovery Software*. ID: G00260499. Stamford, CT: Gartner, Inc. https://www.globanet.com/sites/default/files/resources/Gartner%20Magic%20Quadrant%20for%20eDiscovery%20Software%202014.pdf.

Zierau, Eld, and Sébastien Peyrard. 2016. "Digital Preservation Metadata in a Metadata Ecosystem." In *Digital Preservation Metadata for Practitioners*, 189–211. Springer. https://doi.org/10.1007/978-3-319-43763-7.

Council on Library and Information Resources

1707 L Street NW, Suite 650, Washington, DC 20036-4201
Tel: 202.939.4750 • Web: www.clir.org