

Digital Preservation at Oxford and Cambridge

A collaborative research project to evaluate and provide sustainable recommendations for our digital preservation programmes

Email preservation 2: it is hard, but why?

Posted on **29 March, 2018** by **Sarah**

A post from Sarah (Oxford) with input from Somaya (Cambridge) about the 24 January 2018 DPC event on email archiving from the Task Force on Technical Approaches to Email Archives.

The discussion of the day circulated around what they had learnt during the year of the task force, that personal and public stories are buried in email, considerable amounts of email have been lost over previous decades, that we should be treating email as data (it allows us to understand other datasets), that current approaches to collecting and preserving email don't work as they're not scalable and the need for the integration of artificial intelligence and machine learning (this is already taking place in legal professions with 'predictive coding' and clustering technologies) to address email archives, including natural language processing functions is important.

Back in July, Edith attended the first DPC event on email preservation, presented by the [Task Force on Technical Approaches to Email Archives](#). She blogged about [here](#). In January this year, Somaya and I attended [the second event](#) hosted again by the DPC.

Under the framework of five working groups, this task force has spent 12 months (2017) focused on five separate areas of the final

report, which is due out in around May this year:

- The Why: Overview / Introduction
- The When/Who/Where: Email Lifecycles Perspectives
- The What: The Needs of Researchers
- The How: Technical Approaches and Solutions
- The Path Forward: Sustainability & Community Development

The approach being taken is technical, rather than on policy.

Membership of the task force includes the DPC, representatives from universities and national institutions from around the world and technology companies including Google and Microsoft.

For Chris Prom (from University of Illinois Urbana Champaign, who authored the 2011 [DPC Technology Watch Report on Preserving Email](#)) and Kate Murray's (Library of Congress and contributor to FADGI) presentation about the work they have been doing, you can view their slides [here](#). Until the final report is published, I have been reviewing the [preliminary draft \(of June 2017\)](#) and available documents to help develop my email preservation training course for Oxford staff in April.

So, when it comes to email preservation, most of the tools and discussions focus on processing email archives. Very little of the discussion has to do with the preservation of email archives over time. There's a very good reason for this. Processing email archives is the bottleneck in the process, the point at which most institutions are still stuck at. It is hard to make decisions around preservation, when there is no means for collecting email archives or processing them in a timely manner.

There were many excellent questions and proposed solutions from the [speakers](#) at the January event. Below are some of the major points from the day that have informed my thinking of how to frame training on email preservation:

Why are email archives so hard to process?

1. **They are big.** Few people cull their emails and over time they build up. Reply and 'reply all' functions expand out emails chains and attachments are growing in size and diversity. It takes a donor a while to prepare their email archives, much less for an institution to transfer and process them.
2. **They are full of sensitive information.** Which is hard to find. Many open source [technology assisted review \(TAR\)](#) tools

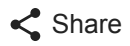
miss sensitive information. Software used for 'predictive coding' and machine learning for reviewing email archives are well out of budget for heritage institutions. Manual review is far too labour intensive.

3. **There is no one tool that can do it all.** Email preservation requires 'tool chaining' in order to transfer, migrate and process email archives. There are a very wide variety of email software programs which in turn create a many different email file format types. Many of the tools used in email archive processing are not compatible with each of the different email file types; this requires a multiple file format migrations to allow for processing. For a list of some of the current available tools, see the Task Force's list [here](#).

What are some of the solutions?

1. **Tool chaining will continue.** It appears for now, tool chaining is here to stay, often mixing proprietary with open source tools to get workflows running smoothly. This means institutions will need to invest in establishing email processing workflows: the software, people who know about how to handle different email formats etc.
2. **What about researchers?** Access to emails is tightly controlled due to sensitivity restraints, but is there space to get researchers to help with the review? If they use the collection for research, could they also be responsible for flagging anything deemed as sensitive? How could this be done ethically?
3. **More automation.** Better tool development to assisted with TAR. Reviewing processes must become more automated if email archives are ever to be processed. The scale of work is increasing and traditional appraisal approaches (handling one document at a time) and record schedules are no longer suitable.
4. **Focus on bit-level preservation first.** Processing of email archives can come later, but preserving it needs to start on transfer. (But we know users want access and our institutions want to provide this access to email archives.)
5. **Perfection is no longer possible.** While archivists would like to be precise, in 'scaling up' email archive processing we need to think about it as 'big data' and take a 'good enough' approach.

SHARE THIS:



This entry was posted in [born-digital](#), [digital preservation](#) by [Sarah](#). Bookmark the [permalink](#) [<http://www.dpoc.ac.uk/2018/03/29/email-pres-2/>] .

About Sarah

Digital Preservation Specialist - Outreach and Training:
Bodleian Libraries, Oxford University

[View all posts by Sarah](#) →

This site uses Akismet to reduce spam. [Learn how your comment data is processed.](#)