

# Digital Preservation at Oxford and Cambridge

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

## Email preservation: How hard can it be?

Posted on **7 July, 2017** by [ehalvarsson](#)

*Policy and Planning Fellow Edith summarises some highlights from the Digital Preservation Coalition's briefing day on email preservation. See the [full schedule of speakers](#) on DPC's website.*

Yesterday Sarah and I attended DPC's briefing day on email preservation at the National Archives (UK) in Kew, London. We were keen to go and hear about latest findings from the [Email Preservation Task Force](#) as Sarah will be developing a course dedicated to email preservation for the DPOC teaching programme. An internal survey circulated to staff in Bodleian Libraries' earlier this year showed a real appetite for learning about email preservation. It is an issue which evidently spans several areas of our organisation.

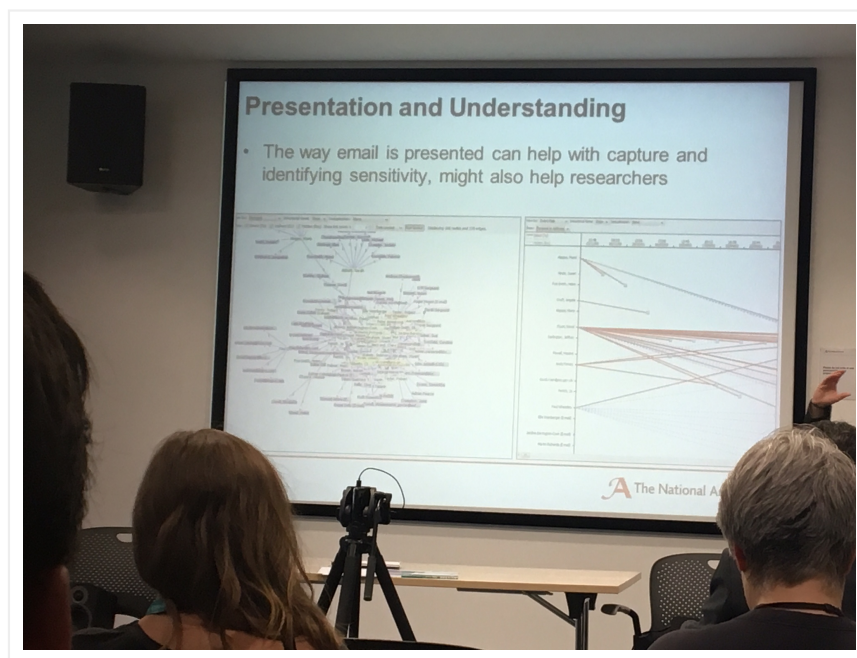
The subheading of the event "How hard can it be?" turned out to be very apt. Before even addressing preservation, we were asked to take a step back and ask ourselves:

*"Do I actually know what email is?"*

As Kate Murray from the Library of Congress put it: "email is an object, several things and a verb". In this sense email has much in common with the World Wide Web, as they are heavily linked and

complex objects. Retention decisions must be made, not only about text content but also about email attachments and external web links. In addition, supporting features (such as instant messaging and calendars) are increasingly integrated into email services and potential candidates for capture.

Thinking about email “as a verb” also highlights that it is a cultural and social practice. Capturing relationships and structures of communication is an additional layer to preserve. Anecdotally, some participants on the Email Preservation day had found that data mining, including the ability to undertake analysis across email archives, is increasingly in demand from historians using big data research techniques.



- Anthea Seles, National Archives (UK), talks about visualisation of email archives.

## What are people doing?

So what are organisations currently doing to preserve email? A strength of the Email Preservation Taskforce’s [new draft report](#) is that it draws together samples of workflows currently in use by other organisations (primarily US based). Additional speakers from Preservica, National Archives and the British Library supplemented these with some local examples from the UK throughout the day.

The talks and [report](#) show that migration is by far the most common approach to email preservation in the institutions consulted. EML and Mbox are the most common formats migrated to. Each have

different approaches to storing either single messages (EML) or aggregating messages in a single database file (Mbox). (However, beware that Mbox is a whole family of formats which have varying documentation levels!)

While some archives choose to ingest Mbox and EML files into their repositories without further processing, others choose to unpack content within these files. Unpacking content provides a mode of displaying emails, as well as the ability to normalise content within them.

The British Library for example have chosen to unpack email files using [Aid4Mail](#), and are attempting to replicate the message hierarchy within a folder structure. Using Aid4Mail, they migrate text from email messages to PDF/A-2b which are displayed alongside folders containing any email attachments. PDF/A-2b can then be validated using vera/PDF or other tools. A CSV manifest is also generated and entered into relevant catalogues. Preservica's out of the box workflow is very similar to the British Library's, although they choose to migrate text content to HTML or UTF-8 encoded text files.

Another tantalising example (which I can imagine will gain more traction in the future) came from one institution who has used [Emulation As A Service](#) to provide access to one of its collections of email. By using an emulation approach it is able to provide access to content within the original operating environment used by the donor of the email archive. This has particular strength in that email attachments, such as images and word processing files, can be viewed on contemporary software (providing licenses can be acquired for the software itself).

Finally, a tool which was considered or already in use by many of the contributors is [ePADD](#). ePADD is an open source tool developed by Stanford University Libraries. It provides functions for processing and appraisal of Mbox files, but also has many interesting features for exploring the social and cultural aspect of email. ePADD can mine emails for subjects such as places, events and people. En masse, these subjects provide researchers with a much richer insight into trends and topics within large email archives. (Tip: why not have a look at the [ePADD discovery module](#) to see it in practice?)

**What do we still need to explore?**

It is encouraging that people are already undertaking preservation of email and that there are workflows out there which other organisations can adopt. However, there are many questions and issues still to explore.

1. Current processes cannot fully capture the interlinked nature of email archives. Questions were raised during the day about the potential of describing archives using linked open data in order to amalgamate separate collections. Email archives may be more valuable to historians as they acquire critical mass
2. Other questions were raised around whether or not archives should also crawl web links within emails. Links to external content may be crucial for understanding the context of a message, but this becomes a very tricky issue if emails are accessioned years after creation. If webpages are crawled and associated with the email message years after it was sent, serious doubt is raised around the reliability of the email as a record
3. The issue of web links also brings up the question of **when** email harvesting should occur. Would it be better if emails were continually harvested to the archive/record management system than waiting until a member of staff leave their position? The good news is that many email providers are increasingly documenting and providing APIs to their services, meaning that the ability to do so may become more feasible in the future
4. As seen in many of the sample workflows from the Email Preservation Task Force report, email files are often migrated multiple times. Especially as ePADD works with Mbox, some organisations end up adding an additional migration step in order to use the tool before normalising to EML. There is currently very little available literature on the impact of migrations, and indeed multiple migrations, on the information content of emails.

### **What can you do now to help?**

So while there are some big technical and philosophical challenges, the good news is that there are things you can do to contribute right now. You can:

- Become a “[Friend of the Email Preservation Task Force](#)” and help them review new reports and outputs
- Contribute your organisation’s workflows to the Email Preservation Task Force report, so that they can be shared

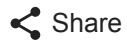
with the community

- Run trial migrations between different email formats such as PST, Mbox and EML and blog about your finding
- Support open source tools such as ePADD through either financial aid or (if you are technically savvy) your time. We rely heavily on these tools and need to work together to make them sustainable!

Overall the Email Preservation day was very inspiring and informative, and I cannot wait to hear more from the Email Preservation Task Force. *Were you also at the event and have some other highlights to add? Please comment below!*

---

#### SHARE THIS:



Share

This entry was posted in [born-digital](#), [digital lifecycle](#) by [ehalvarsson](#). Bookmark the [permalink](#) [<http://www.dpoc.ac.uk/2017/07/07/email-preservation-how-hard-can-it-be/>].

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