Collecting emails

Challenges and complexities
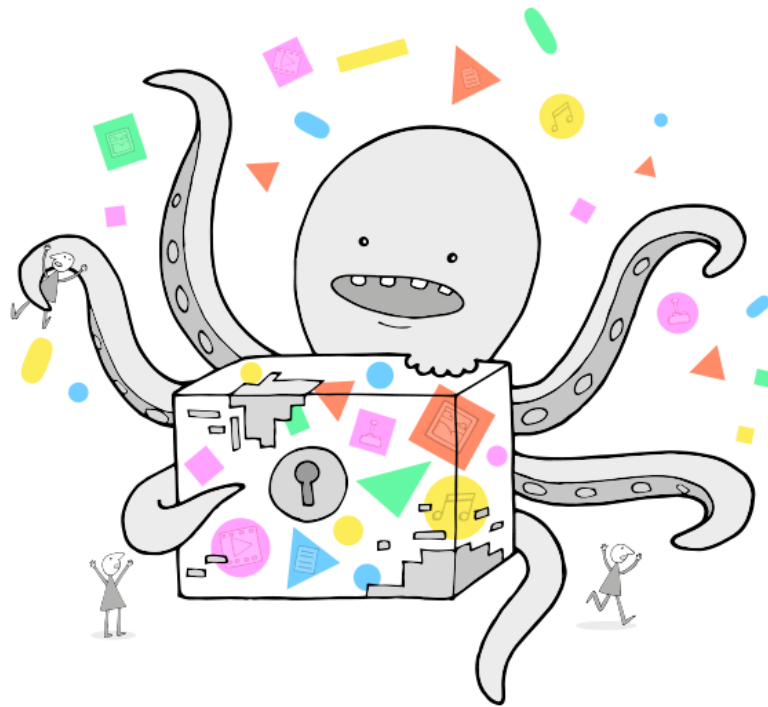
Matthew Burgess

Digital Collections Analyst

State Library of New South Wales
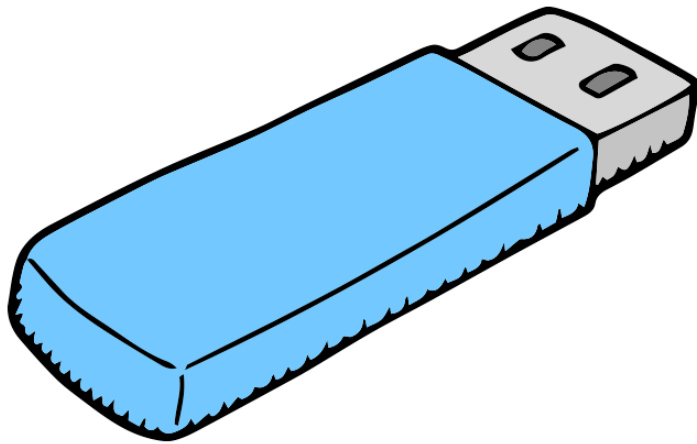
@matthewpburgess

Workflow

Current context for emails at the State Library of New South Wales

- Lack of defined workflow for preservation and access does not impact decisions for acquisition

- Already receiving collections with email components

- Highlighting potential future workflow options encourages staff to acquire

- Multiple solutions in development or research, nothing built in with existing system architecture

Acquisition

- Providing instructions to collection owner
    - →Differs based on email provider/software
    - →General rule is to create a 'folder' of selected emails for export, including both sent and received emails where possible
    - →Impacted by owner's computer literacy
- Acquiring a copy from the owner
    - →File format differs based on email provider/software
    - →Technical limitations of owner's computer for export can have a big impact
- Standard born-digital procedure to 'bag' files on network storage

Valuation

- Digital material requires valuation for inclusion in the collection

- How do you value an email?

  →By size? What about duplication?

  →By content? How do you generate and provide a summary? What information is important?

- Valuation requires access for determination

  →Utilise original software for browsing?

  →Use Forensic Toolkit?

  →Use ePADD?

Sensitivity

- Emails contain a lot of personally identifiable information (PII) in various places, not just the body text

- The content of an email may be sensitive

- Ability to identify and redact PII impacted by file format

- Identification is not as much as a challenge as redaction

Processing

- How do we process email for preservation and access?

  →Converting PST to MBOX?

  →Extract individual emails in PDF or EML format?

  →Analysis in ePADD for archival description?

  →Can we process emails in PDF format in the same way?

- What information should be stored in the catalogue vs preservation system?

Preservation

- Collection Preservation Policy determines preservation of the original file, regardless of format

- Consideration for migration to MBOX for preservation and access where appropriate (e.g. for use in ePADD)

- Established workflows for ingestion and preservation for other file formats, untested for emails

- Bit level preservation is not a challenge

Access

- Plans for a dedicated, restricted computer for accessing digital collections in the reading room

- Can we provide access online with integrated viewers as part of our Collection Experience Program?

- How do researchers want to engage with emails?

  →Emulated in the original computer environment?

  →In depth analysis, eg ePADD, FTK?

References and further reading

- SLNSW Collection Experience Program, https://collection.sl.nsw.gov.au/about

- SLNSW Collection Preservation Policy 2021, https://www.sl.nsw.gov.au/sites/default/files/collection_preservation_policy.pdf

- Australasia Preserves, COVID-19 collecting email meetup, https://www.australasiapreserves.org/2020/05/covid-19-collecting-meetup.html

- Review, Appraisal, and Triage of Mail, https://ratom.web.unc.edu/

- Task Force on Technical Approaches for Email Archives. 2020. "Requirements for Archiving Email Using PDF (Draft)." Accessed 27 May 2021 https://docs.google.com/document/d/1JNij3M2UK2XUdsnRg4j9j33XPaKj5-JSvijwpqlwbRo/edit?usp=sharing

Title slide image adapted from pixabay, https://pixabay.com/illustrations/contact-visit-letters-email-mail-2794680/.

All other images from the Digital Preservation Business Case Toolkit, http://wiki.dpconline.org/, CC BY-NC 3.0.