eMail archiving @ Public Record Office Victoria

Andrew Waugh

Public Record Office Victoria

VICTORIA State Government

Public Record Office Victoria (PROV)

- The archival authority for the State of Victoria
- In Australia States perform most of the internal government functions (police, health, education…)
- PROV sets record standards for State government agencies
- PROV receives, archives, & provides access to key State records no longer in active government use
- *We are not a collecting archive*

Public Record Office Victoria

VICTORIA State Government

Why email?

The smoking gun is always in the email

- Work is done in email
- The final result is saved to formal record systems
- Generalises to modern collaborative environments – Teams, Snapchat

Our goal:

- Capture the email that documents key government decision making
- The challenge is scale: 56,000 employees in central VPS (+ teachers, hospitals, higher ed)
- Even a Capstone approach still results in large numbers of staff
- Our implicit goal: to cull most government email

Our (ongoing) email project

What we've done:

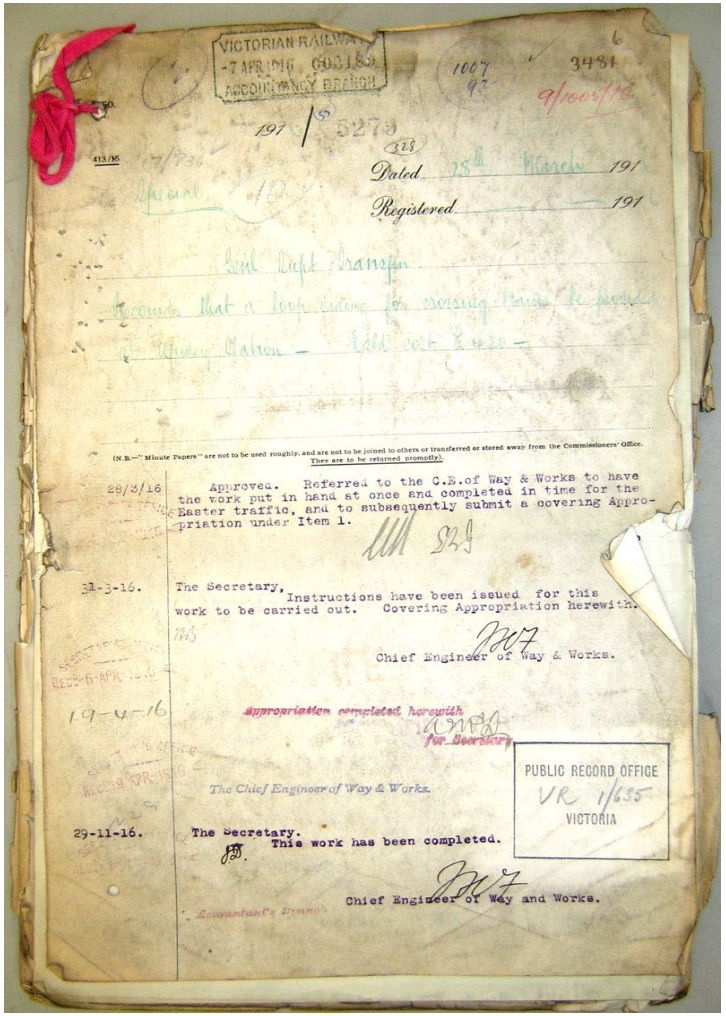
- Tests and pilots to understand the issues & feasibility
- Obtained 2 years of our own email as a test set (~1.2 million emails with 3-4000 emails per account)
- Used an eDiscovery tool (NUIX) to select & cull emails
- Investigated culling criteria
- Prototype migration & ingest workflow

The project is ongoing, but want to share what we have learnt so far

Idea 1: How do you think of email?

As individual structured email accounts or a shared body of email with different views on it?

- Archival principle of original order, but computers have multiple orderings
- Account view is common – reflecting focus on individuals & original order
- Shared body of email is useful with a set of related accounts
- Allows deduplication (40% reduction) & restoration of missing emails
- Can still present as individual accounts



Idea 2: Positive and negative appraisal

Corporate email has email with many purposes:

- Personal (individual)
- Personal (corporate e.g. HR)
- Work (social)
- Work (administration)
- Work (process)
- Work (substantive)

As a government archive, only the last is of value

Our goal is to use automated tools to cull the emails (& especially eliminate personal information)

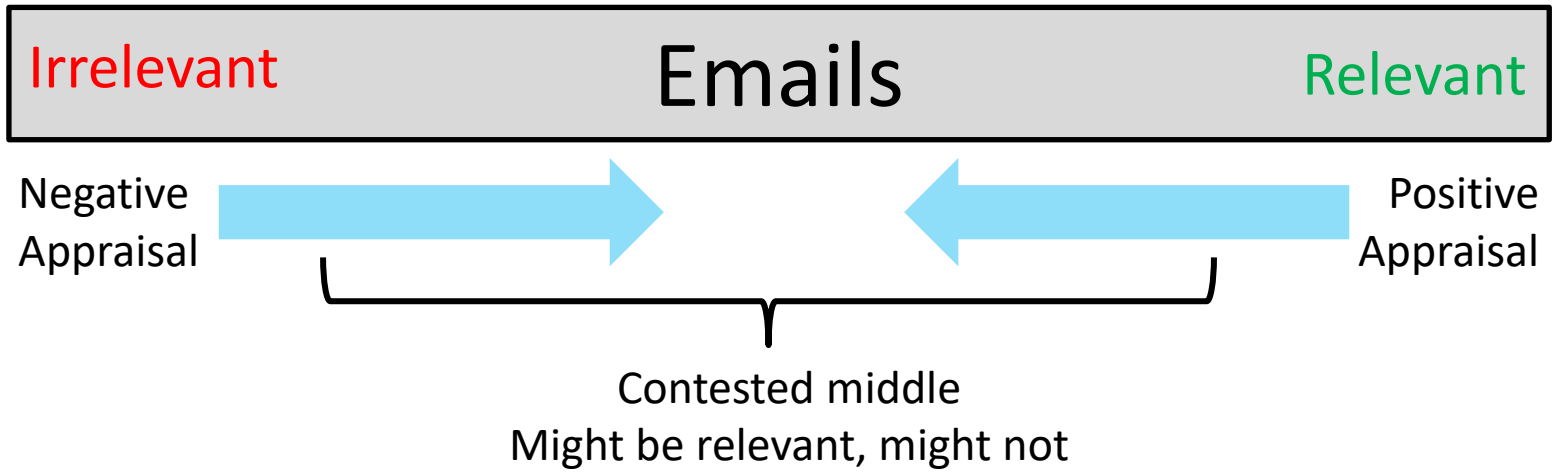
Test: manual trial using subject & sender tags based on one year

- Threaded to reduce replication (9347 threads)
- Manual inspection of subject & sender to infer value (aided by it being PROV's email – we are familiar with our business)
- 69% judged to be ephemeral or non permanent (general to all agencies)
- 4% judged to be ephemeral or non permanent (PROV specific)
- Produced a list of generic terms for culling

Positive and negative appraisal

Two ways of thinking about appraisal…

- Positive appraisal: select the emails we are interested in (substantive work) & discard the rest
- Negative appraisal: select the emails we are NOT interested in & keep the rest



Benefits of negative appraisal

Positive appraisal has lot of appeal, but…

- Positive appraisal means selecting on characteristics unique to the agency (key work emails are specific to an agency's unique business)
- Negative appraisal selects on characteristics of email that are more likely shared between agencies (HR, general admin, personal)

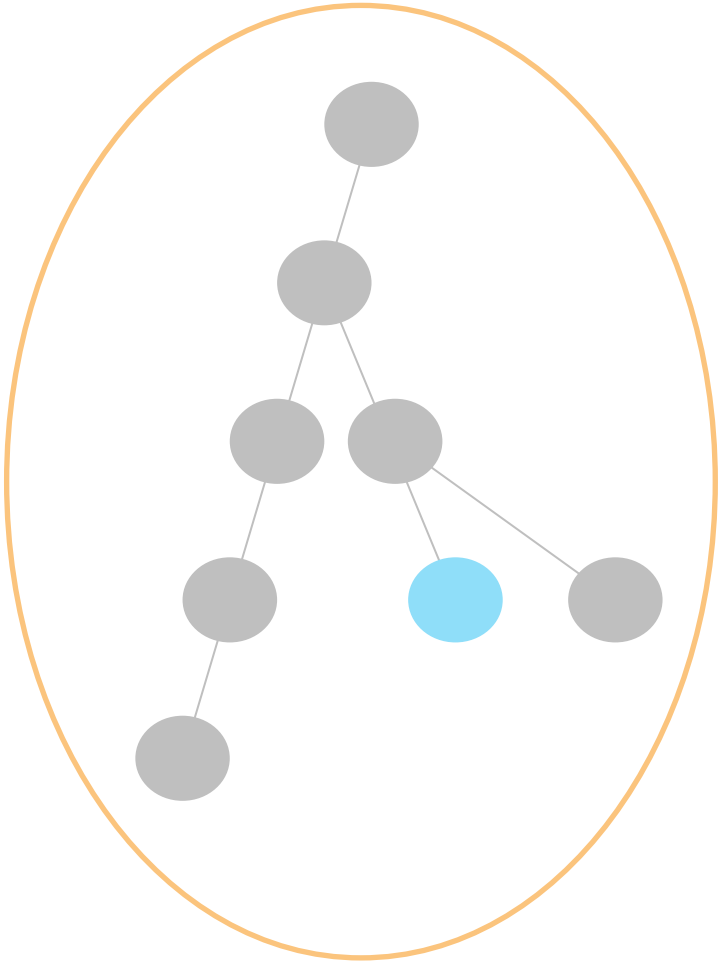
Negative appraisal:

- Greater scope for generalising automated tools across agencies and across jurisdictions
- Particularly valuable in building training sets for AI tools

Idea 3: Threading email

Using ‘Reply’ and ‘Reply-all’ automatically links (threads) the reply to the original email

- Brilliant for researchers
  - Find one relevant email, read the thread of related emails
  - Reduces clutter by presenting threads, not emails
- Brilliant for appraisal
  - Reduces number of decisions
  - Increases information (threads not individual emails) available to automated tools



Threading results

Positives

- Impressive reduction in clutter when viewing email collections (60% reduction of deduped emails – thread length average 2.4)
- Even manual appraisal felt more achievable
- Useful even if only considering one account
- Cheap to implement

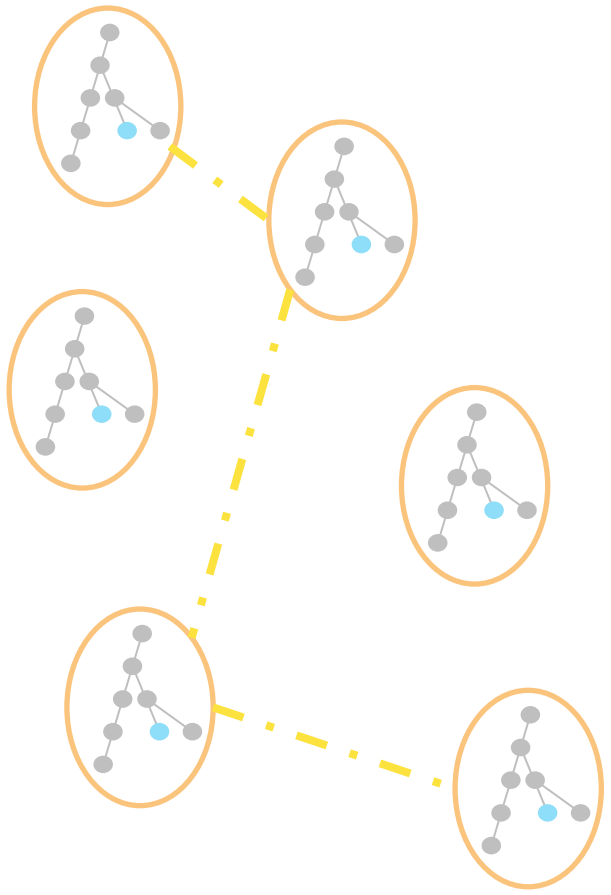
Negatives

- Thread drift
- Thread breaks
- Technical harder than it should be because not all email systems follow the standards
- Length of threads does not indicate importance! (Our longest thread was 335 emails about replacement office chairs)

Idea 4: Super threading

Use AI tools to overcome thread breaks – what thread came before/after this thread

- Uses the additional information available in a thread of information
- Researcher finds one email, reads thread, finds related threads…
- Might even involve different groups of people
- For the future!



Idea 5: eDiscovery tools

eDiscovery tools are commercial products used in legal work to examine collections of data such as email

- Discovery in civil cases
- Investigations
- Similar in concept to ePADD

Our pilot used an eDiscovery tool (NUIX) to process the email collection

- Deduplicate
- Thread
- Appraisal
- Migration to archival format
- Visualisation tools

Observations on eDiscovery tools

Positive:

- Worked
- Very powerful tool ranging from selecting on metadata to visualisation and simple AI
- Supported
- Could process multiple accounts at once
- Manual selection model, but could use batch processing

Ultimately, too expensive for us

Negative:

- Expensive (ongoing license)
- VERY hard to use (vendor support/training required)
- Requires a VERY high end computer
- Do not use virtual (cloud) computers due to transfer rates
- Logic (e.g. deduplication, threading) opaque and could not be tuned
- Fine grained selection of questionable value – we were getting good results with simple culling techniques. Need ML or AI techniques to be worth more

Idea 6: The challenge of test data

Need real appropriate email test data to carry out tests (and ultimately to train AIs)

- How representative is the test data in our domain?
- How portable is our domain?
- Privacy and sensitivity challenges
- Can others reproduce our results?

Our test data - PROV emails for two years – all staff, all emails

- Clearly sensitive, both from an organisational and personal perspective
- Required clear communication with staff (and management) about what we were doing
- Tests relating to email content limited to tester's own email
- Service provider concerns
- Absolutely cannot share our test data

Take away messages

- The purpose of your archive will affect how you think about email archiving
- PROV thinks of email as a collaborative workspace presented as structured individual accounts, leading to thoughts of deduplication and restoration
- Negative appraisal has the possibility of selectively culling the private and junk while producing a portable tool
- The use of threading to concentrate emails to assist in processing and research.
- The possibility of super-threading
- eDiscovery systems are powerful, but expensive and not a perfect fit for archival purposes
- The challenge of obtaining test data to build email processing systems

Thank you

VPRS 12800/P0001, ILLUMINATIONS FLINDERS STREET CENTENARY 1954

Public Record Office Victoria

VICTORIA State Government