Developing a practical end-to-end solution

for managing email in Victorian Government

Evanthia Samaras, Public Record Office Victoria

Agenda

1. Introduction

About the presenter, PROV and VERS

2. Background to the email pilot project

Outline of problem, previous email PoC project work

3. Email pilot project overview

Objective, research questions, outline of tasks to be undertaken

4. Outcomes

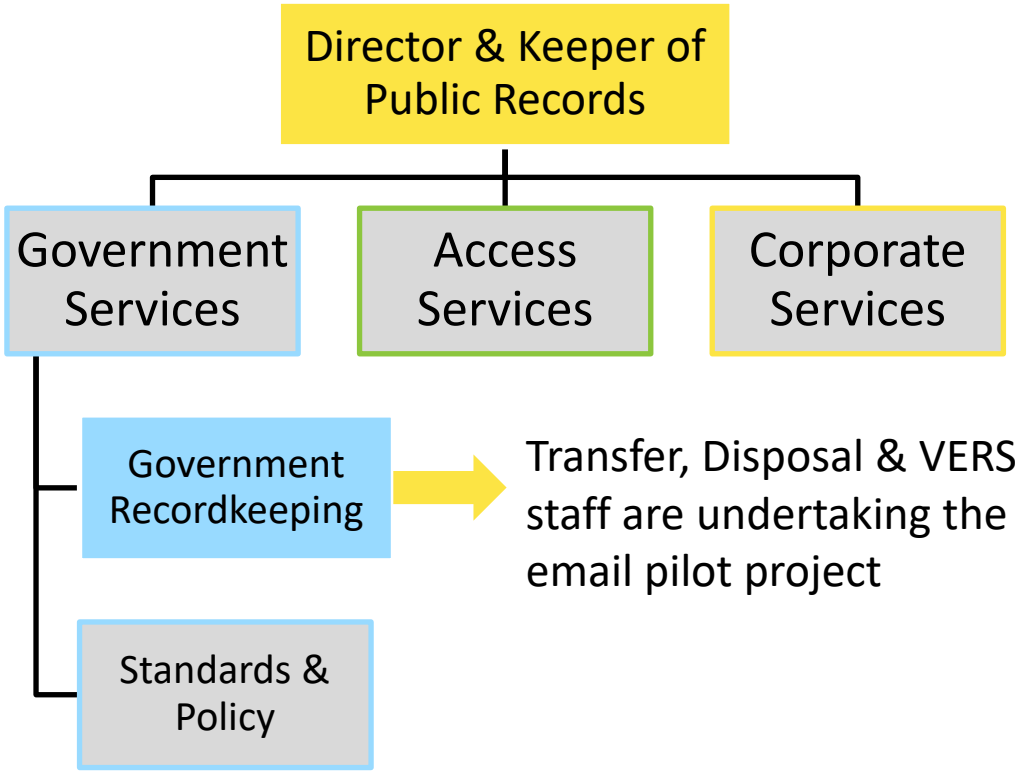
Potential long-term outcomes for the pilot project

Introduction

About Public Record Office Victoria (PROV)

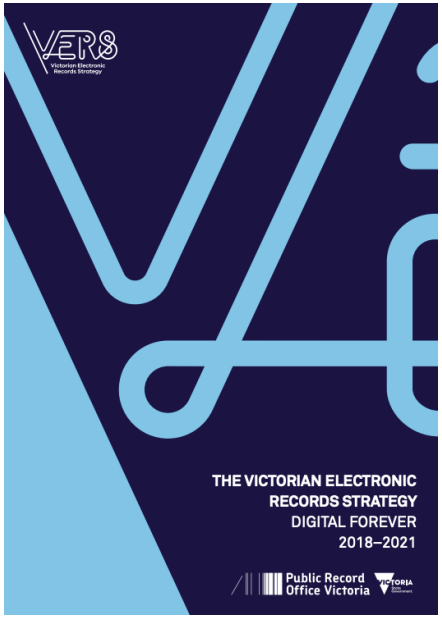
- PROV is the Victorian State Government archive
- Established in 1973 with records dating back to 1830s
- Located at the Victorian Archives Centre in North Melbourne
- We've been actively transferring digital records since our first Digital Archive was implemented in the early 2000s

PROV organisational structure and teams



Victorian Electronic Records Strategy

Victorian Electronic Records Strategy



See the Strategy at: https://prov.vic.gov.au/recordkeeping-government/vers

Background to the email pilot project

The email problem in Victorian Government

Since the late 1990's the Victorian Government (VG) has used IBM's Lotus Notes (LN) email application

Over 20 years of routine backup by the government's IT provider, CenITex, has resulted in a huge build-up of email storage

Large build-up of emails

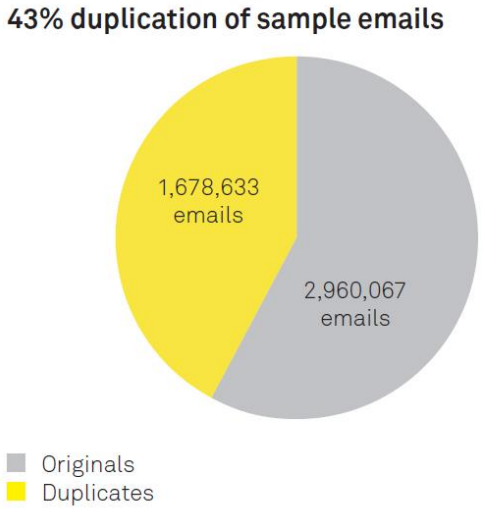
An audit in 2014 revealed that CenITex had over 67,000 linear tape-open (LTO) magnetic tapes in storage and over 28 petabytes of online storage

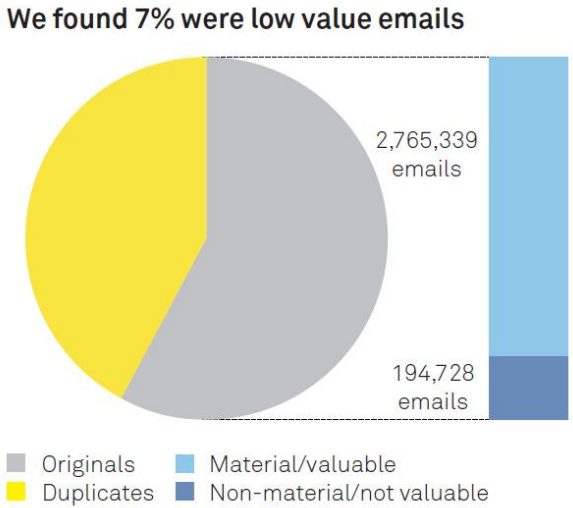
VG is currently in the process of moving to MS Outlook and Office 365 — What will happen to all the backed-up LN emails?

Previous email project work

In 2017/18 PROV undertook a proof of concept (PoC) project to test a commercial eDiscovery tool on a set of emails.

This project focused on disposal and de-duplication functions.





Email pilot project overview

Objective

The new project to be undertaken during 2020 will aim to develop an end-to-end machine-assisted solution to appropriately appraise, dispose (and potentially transfer) emails

The project will analyse a corpus of online PROV emails (2016 to present, approx. 500 GB). It will not address the emails currently held on LTO tapes.

It will conduct further tests with the eDiscovery tool we used in the PoC and also test available open source machine learning (ML) / natural language processing (NLP) tools.

Approach

Test available commercial and open source tool functionalities across a large collection of emails.

If sufficient appraisal outcomes are achieved…

Then we will proceed to transfer the permanent value emails to the PROV Digital Archive

Questions to explore

Using commercial e-discovery tools and/or open-source tools…

- Which have the means to convert proprietary IBM NSF email format?
- Which have the potential to clean, analyse and appraise emails?
- What de-duplication functions are available to use and how do these compare?
- Can analysis and appraisal processes be conducted across email threads (i.e. conversations) as opposed to individual emails?
- Can appropriate rules be applied to identify non-record, long-term temporary and permanent value emails across a collection?

Project tasks

Step 01
Setup Environment

Step 02
Load emails

Step 03
Identify and source tools

Step 04
Convert emails

Step 05
De-Duplicate

Step 06
Remove non-records

Step 07
Appraise and dispose

Step 08
Sensitivity review

Step 09
Arrange and describe series

Step 10
Prepare VEOs

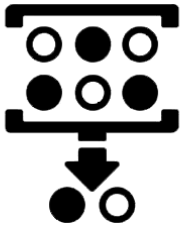
Step 11
Ingest

Step 12
Access sign off

Initial testing with e-discovery tool

Recent tests with 1 x year of PROV emails:

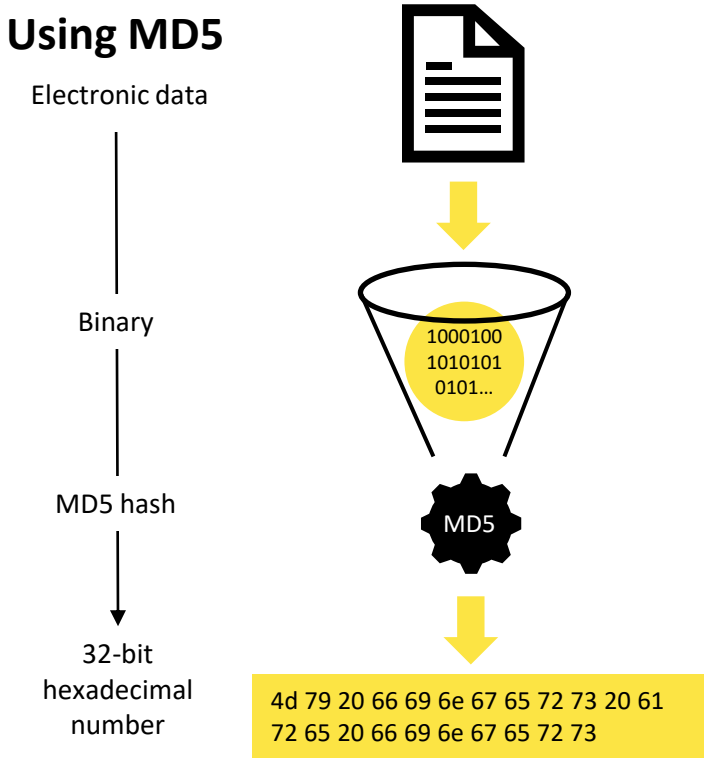
- 79 x accounts from 2018
- Processed original NSF file format
- 249,683 emails
- 57.4 GB



'Simple' de-duplication function performed

- 142,634 emails after de-dupe
- 57% of original amount remaining after de-dupe
- 4 hours processing time (using under-configured system)

How de-duplication works with the tool



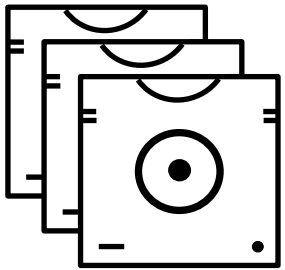
The tool we used has a custom proprietary MD5 hash solution for email that uses:

- Subject
- From
- To
- CC
- Email body (text tokenised so whitespace and irrelevant characters are removed)
- Binary streams of attachments

Outcomes

Potential project outcomes

If we can achieve a practical and sufficient end-to-end machine-assisted solution to appropriately appraise and dispose a large collection of emails, we can…



Move away from LTO as the email archives source

Get departments to work with PROV to roll out email appraisal, disposal and transfer

Convert permanent emails into VERS Encapsulated Objects (VEOs)

Transfer permanent emails to PROV's digital archive for ongoing access and preservation


Thank you
Any questions?