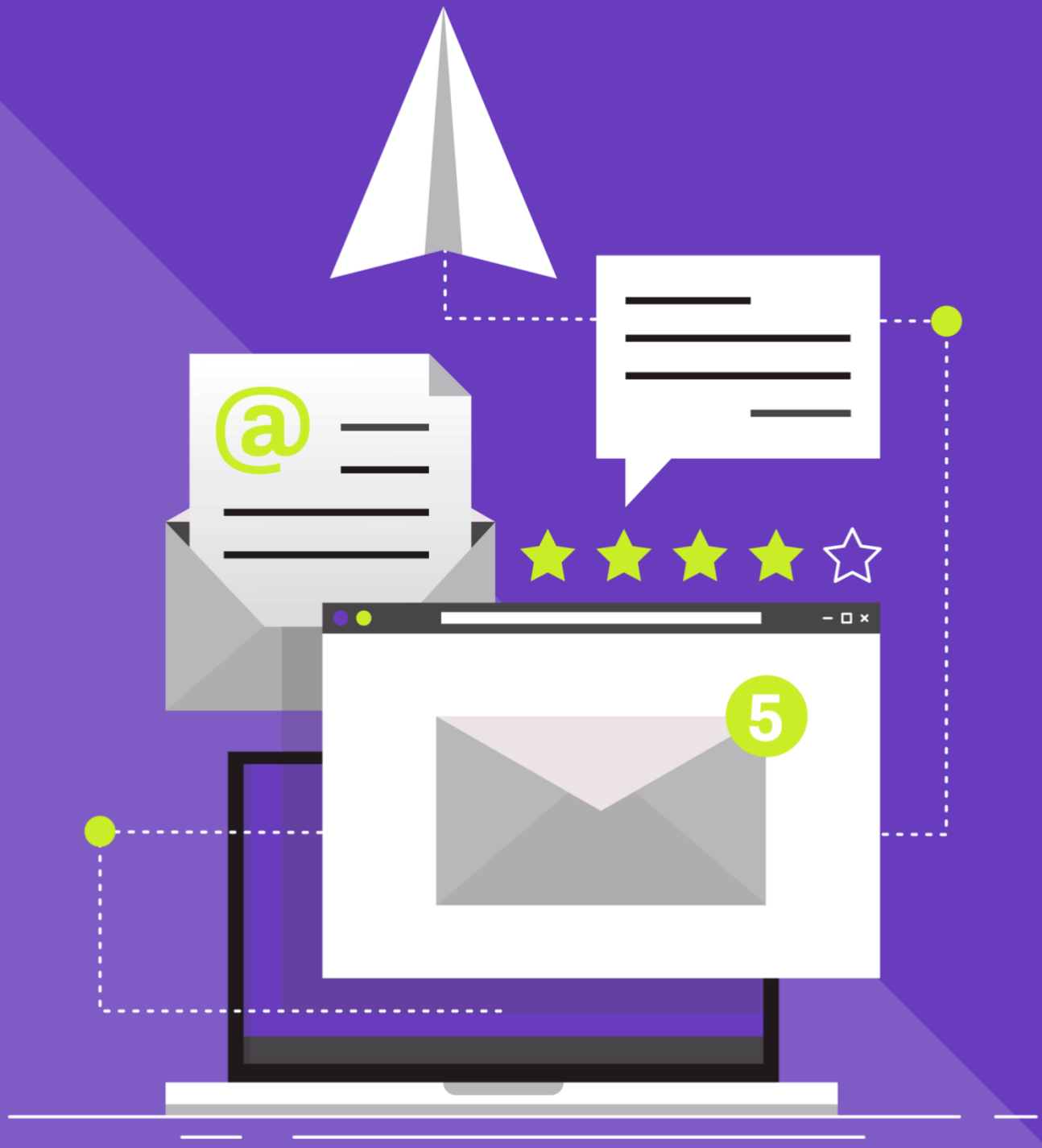
Public Record Office Victoria

Stage 2 Email Appraisal, Disposal and Preservation Project 2019-2020

Summary Report



VERS Victorian Electronic Records Strategy

Public Record Office Victoria

VICTORIA State Government

Document Information	
Title	*Stage 2 Email Appraisal, Disposal and Preservation Project 2019-2020: Summary Report*
Version	1.0
Approved by	Justine Heazlewood
Date	5 November 2020
Business Owner	David Brown
Authors	Evanthia Samaras, Julie McCormack and Andrew Waugh
Classification	Public

The project

Since the late 1990s the Victorian Government has used the IBM Lotus Notes email application as a primary communication medium. Key actions and decisions of public officers are captured in the email, forming a primary repository of government records. In its current proprietary format however, its value as an information source cannot be fully realised, posing not only a risk to informed decision making for current administration, but a potential gap in the documented memory of Victoria. Routine back-up has resulted in a backlog of over 28 petabytes of content, adding to management costs and compromising government accountability and transparency.

To address this pressing recordkeeping issue, Public Record Office Victoria (PROV) in collaboration with our government partners, is committed to investigating methods to preserve Lotus Notes emails for the benefit of government and the public good[1]. This work aims to meet Goal 2 of the *Victorian Electronic Records Strategy 2018-2021*: "Preserved and accessible digital records of continuing value: Identifying barriers to transfer and developing solutions"[2].

We conducted a Stage 1 project in 2017-18, which tested an e-discovery tool on a sample set of Lotus Notes emails. The project demonstrated that up to 50% of the email backlog could be culled efficiently through the targeted application of an e-discovery tool. The tool could also add metadata to emails to assist classification and processing for preservation and discovery.

During 2019-20, a Stage 2 project was conducted to build on the findings of Stage 1. The e-discovery tool was used again, supplemented by open source tools.

The Stage 2 project was led by staff at PROV with assistance from:

- CenITex Victorian Government information technology service provider
- Nuix e-Discovery software company.

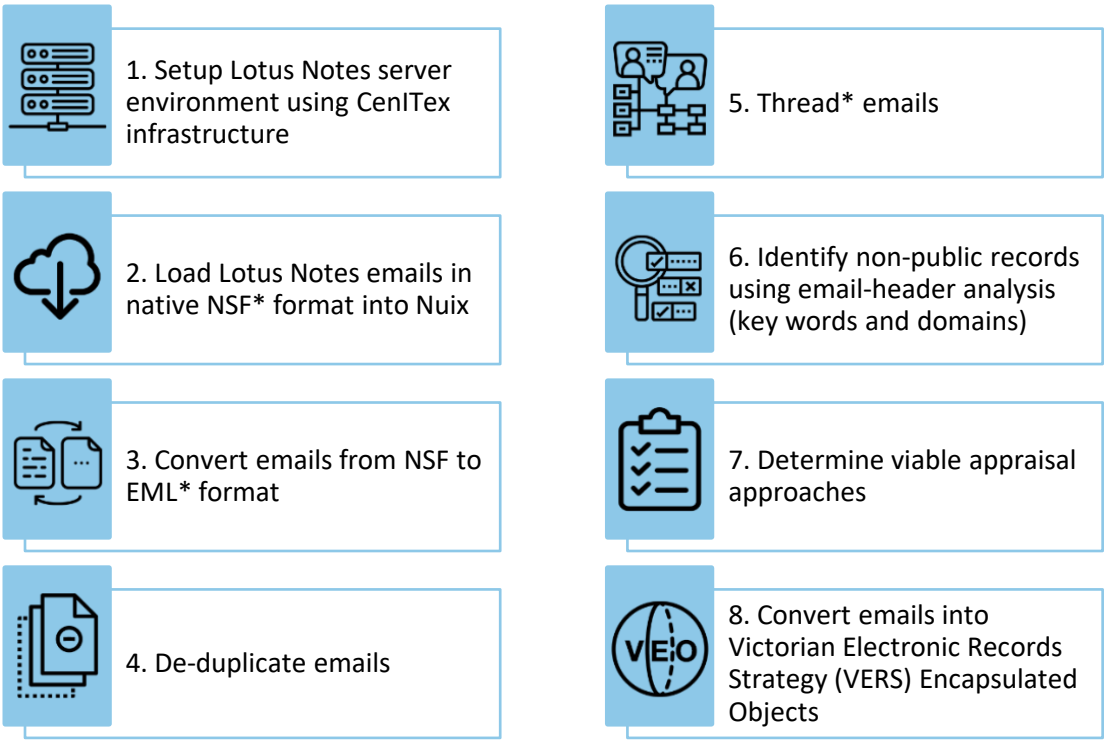
The Stage 2 project enabled PROV to develop a deeper understanding of the Lotus Notes Storage Format (NSF) and its challenges. This knowledge will inform our advice to government on the management of past, current and future email implementations, including Microsoft 365.

The project confirmed that NSF is not a sustainable format for email records, supporting the need for Victorian Government to start managing its backlog of Lotus Notes emails before they become completely obsolete.

[1] The international recordkeeping community is yet to develop a scalable solution for email appraisal and preservation applicable to organisational Lotus Notes email accounts.

[2] See the Strategy at: https://prov.vic.gov.au/sites/default/files/files/VERS/VERS_Digital_Forever_2018-2021.pdf

Project tasks



*See definitions below:

EML is a common file format used by many email clients including Microsoft Outlook and Apple Mail. It contains the content of the message, along with the subject, sender, recipient(s), and date of the message. EML files may also store one or more email attachments, which are files sent with the message. EML files can be opened in a variety of email programs as well as text editors.

NSF stands for Notes Storage Format. NSF files are supported by IBM Domino Server as well as IBM Lotus Notes. An NSF file is capable of holding crucial databases of IBM Lotus Notes including emails, contacts, design information, user data and appointment.

Thread refers to a single email conversation that starts with an original email, (the beginning of the conversation), and includes all of the subsequent replies (and, optionally, forwards) pertaining to that original email.

Email sample

CenITex provided PROV with a copy of all available *online* PROV staff email[3] accounts between 2016 and 2019. Pre 2016 emails stored in Linear Tape-Open (LTO) format were not in scope, due to the extra time and resources required to retrieve and access the emails from LTO.

1.2 million emails

Complete email accumulations from 2016 to 2018

Partial accumulations from 2015 and 2019

NSF files for each user for each year

Process

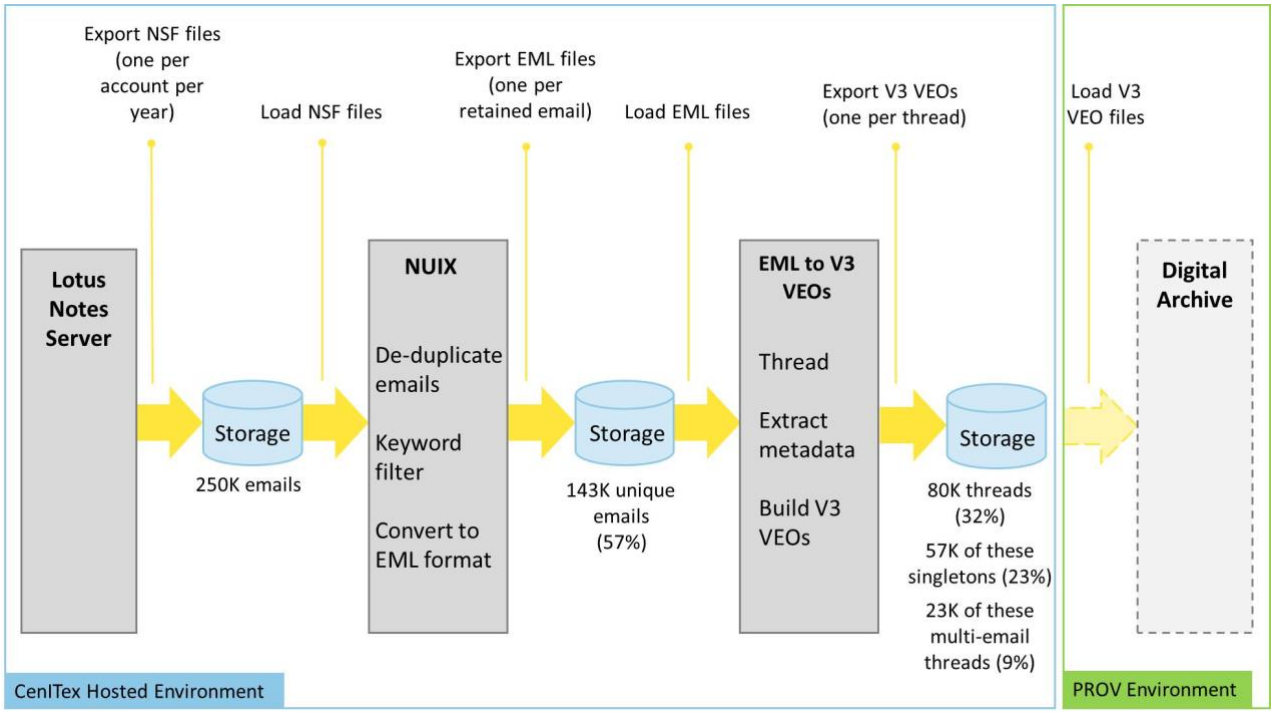


Figure 1 Stage 2 process

The process began with setting up the CenITex virtual machine environment and adding a Lotus Notes Server. Then the NSF files were loaded and imported into the Nuix e-discovery tool for processing. A small number of files of one user's email were then exported as EML format for threading task tests, which were achieved using bespoke software created by PROV. We also did some threading tests using the built-in threading function in Nuix. Finally, we undertook tests creating version 3 VEOs using individual email records and threaded email record approaches.

The dashed lined section at the far right of the diagram represents the Digital Archive ingest task that was intended to be included in the process but was not undertaken due to delays in the implementation of PROV's new Digital Archive system.

[3] PROV email was chosen in order to simplify the management of privacy and other information management regulatory frameworks.

Key findings

Described below are the key technical, recordkeeping and information management findings of the Stage 2 project.

Technical findings

Lotus Notes' NSF format is an unsuitable format for long-term access and preservation of email.

- The very qualities that made Lotus Notes attractive for government email (mainly its security attributes) make NSF a difficult format to manage and access long term.
- NSF is an undocumented, proprietary, binary format that relies upon the Lotus Notes client to access the data.
- Availability and access to email processing tools that support NSF is limited.
- Similar undocumented, proprietary, binary formats should be avoided by the public sector.

It is expensive to process and analyse emails.

- Available software that can process and convert NSF format is generally commercial with licence fees. Vendor support also generally has a fee.
- A powerful processing environment is required due to the volume of emails and their complexity.

It is challenging to transfer the large amount of data required to and from hosted cloud environments.

- Cloud environments are less usable than processing on desktop machines. This is due to the amount of data that needs to be transferred to and from the computing environments. The security requirements imposed on access can make using the cloud environments difficult.

The lack of sophisticated technologies limits appraisal processing.

- Access to high performance Machine Learning and/or Artificial Intelligence software is not practically or financially achievable for Victorian Government until the technology becomes more readily available and affordable to the broader market.
- This reduces the ability to conduct appraisal across unstructured email accumulations to more precisely identify public records of continuing value.

Recordkeeping and information management findings

Threading contextualises emails, making them more understandable and usable records.

- Threading echoes traditional filing of paper records that links related records together, allowing easy reading by humans, in particular it enables the reader to follow the chronology of events, enhancing its evidential value.
- Managing threads rather than individual emails significantly reduces the volume of records. This aids access and simplifies and reduces the cost of archival management.
- Threading email conversations improves understanding of organisational context.
- A thread provides more enhanced documentary evidence for management and discovery than an individual email removed from its communication context.

Working with consolidated email for the organisation is better than emails of individuals.

- If one participant in an email exchange keeps a copy, the email survives.
- Threads remain intact as participants of the email exchanges change.
- Selection focus is on the email/thread subject, not on individual participants.
- De-duplicating consolidated email accumulations provides significant reduction in email quantities – around 40% of emails were found to be duplicates.

Culling of email to remove non-public records can be achieved using domain name and header keyword analysis.

- Non-public record emails can be identified by applying a sample of domain names and an ontology of common terms across email headers.
- Analysis of the body content may not be required to identify non-public records[4].

It is very difficult to apply functional appraisal and classification to unstructured content.

- In the absence of machine learning and/or artificial intelligence technology, classification is very challenging and requires a high level of manual intervention and review.
- Established archival techniques for managing unstructured hardcopy correspondence (i.e. removing easily identifiable temporary value correspondence such as those related to finance, HR and facilities records), may prove effective for managing and preserving email until sophisticated ML and/or AI technology becomes more readily available to Victorian Government.

Email accumulations should be appraised according to organisational recordkeeping systems.

- Email cannot be appraised effectively in isolation. The functions of the organisation and its recordkeeping practices influence whether organisational email may be of continuing value and should be retained permanently.

IT management regimes can disturb the original order of the organisational email.

- Older email content and ceased employees' email accounts are removed from online storage to tape, artificially separating email accounts from their original organisational context.
- IT specialists should collaborate and partner with recordkeeping specialists on back-up planning to reduce the impact of backup on records' integrity and context.
- Disclaimers to explain the disturbance to the original order of existing email should be made available for future users of the records.

The scale of email means it is impossible to account for every email.

- It is difficult to measure and validate the accumulation of email (including confirming the correct number of email accounts) extracted and processed.
- Approaches such as removal on non-public records and confidential emails cannot be guaranteed to be 100 per cent accurate.

Next steps

Overall, the project has provided some promising results and viable practical options to help manage Victorian Government Lotus Notes email accumulations into the future. Over the coming year, the project team will commence Stage 3 project planning. It is anticipated that Stage 3 may include the following work:

- Test available tools to compare against Nuix for de-duplication and NSF conversion tasks.
- Determine ways to effectively identify and protect confidential emails.
- Continue header analysis by expanding terms and develop a list of domains for the task of identifying non-public records.
- Improve threading approaches and software scripts (and potentially explore how to tie related threads together).
- Continue VEO creation tests and determine appropriate construction arrangement for email record series.
- Arrange and describe the PROV email series and convert into VEOs for ingest into the new PROV Digital Archive.

[4] This reduces security concerns of accessing email content.