

Gmail: Harvesting and Ingesting Executive Director Data

Elizabeth Perkes
Utah State Archives

GroupWise

- ▶ For 20 years, Utah used GroupWise as its enterprise email system
- ▶ GroupWise export options were:
 - Individual emails, saved as text or .eml
 - Whole accounts, saved as XML, accessible via Nexic client
- ▶ Limited searching options, bulk exports done on the backend by IT
 - Public records requests very time-intensive, and expensive

GroupWise

- ▶ Archives has email from two accounts for people who left prior to Gmail conversion:
 - Budget officer of Archives
 - A few dozen emails
 - Former state CIO who left after a data breach
 - Tens of thousands of emails
- ▶ Nexic data not very well self-described
 - Relies on local executable that isn't being updated with OS changes
 - Folders not named in meaningful way, unknown XML structure
 - Client is user-friendly

Show me

- ▶ <http://archives.utah.gov/axaem/movies/Nexic/Nexic.htm>

Gmail

- ▶ In 2012, Utah transitioned to Gmail
 - Funding for this change was available as it impacted databases integrated with email
 - Archives was able to connect to the Gmail API with its AXAEM system
 - Used this feature to send emails from an existing Gmail account via the AXAEM interface, impacting:
 - Records officer online training/certification
 - Patron requests for records, ordering boxes from storage

Gmail

▶ Apple Valley, UT

- Used Gmail
- ISP sold their domain name, no access to email
- Called Archives for help
- Archives asked APPX how to download this email
- APPX created simple interface using existing Gmail API
- Interface now used regularly:
 - By Archives, to harvest executive director data
 - By DTS, to respond to litigation and security investigations; or agencies answering public records requests
 - By agencies, because they want an easy way to move data offline, especially those leaving state employment, or share data with third parties

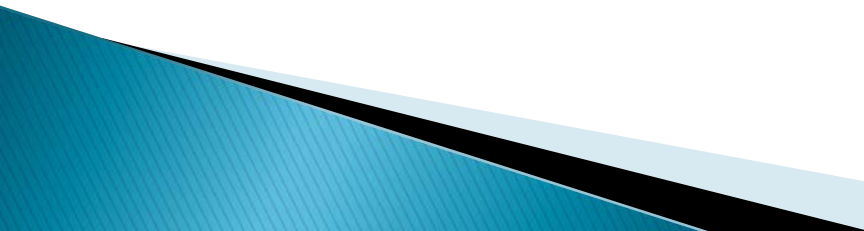
Gmail

- ▶ Multiple labels can be assigned to the same email, different from concept of “folders”
 - Search email in Gmail using advanced search
 - Select hits
 - Apply a label to the hits
- ▶ Log into AXAEM
 - Provide Gmail account name and password
 - Click “Extract Contents”
 - Indicate location where email is to be saved
 - Select labels whose contents you want to export
 - Click “OK”

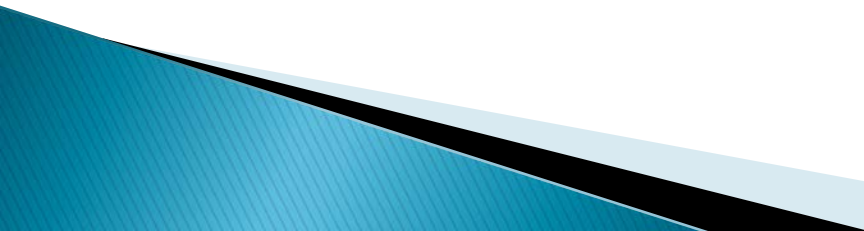
Show me

- ▶ <http://archives.utah.gov/axaem/movies/gmail/gmail.htm>

EML as Preservation Copy

- ▶ Stored as plain text
 - ▶ Metadata easy to extract
 - ▶ Desktop email clients know how to render it, make attachments viewable
 - ▶ Attachments encoded as base64, which can be transformed to a binary and stored separately if desired, or migrated forward
 - ▶ Easy to de-accession if content not preservation-worthy
- 

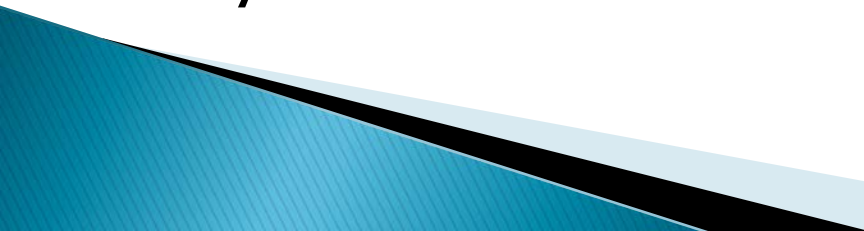
Valuable email saved

- ▶ Public Safety
 - ▶ Transportation
 - ▶ Facilities Construction & Management
 - Non-director, 30-year employee asked for copies of his email
 - Found conversations with Capitol Preservation architect, who left long ago without email being saved
 - Found minutes to lots of meetings, plenty of value in his email account, though he wasn't a director
- 

Problems with Directors' email

- ▶ Used Google Docs instead of attachments
 - Have to read each email to know if a link is there
 - Have to have account/password still active in Gmail to access
 - Once you download the file, how do you associate it with the email, stored in context?
- ▶ Outgoing directors wiped email accounts, some forgot to do so with sent mail 😊
- ▶ Inbox keeps filling up with messages even after they left, hard to know termination date
- ▶ Sent mail filled with auto-replies ☹️

Appraisal & Non-Public Data

- ▶ With tens of thousands of emails to sift through, how can we weed accounts?
 - ▶ Agencies could apply labels indicating retention and access restrictions
 - ▶ Need appraisal interface for exported email
 - ▶ To create a redacted copy, need a way to transform to PDF and use Acrobat's features
 - ▶ Need way to associate redacted copy with original during ingest into preservation system
- 

How to Ingest Email

- ▶ Our ingest procedure is this:
 - Use BagIt to capture files with manifest and checksums, write to M-disc.
 - Upload bag to AXAEM, where checksum is verified valid
 - Metadata from records extracted and written to database

Show me

- ▶ <http://archives.utah.gov/axaem/movies/ingest-email/ingest-email.htm>
 - Demo omits BagIt step in the interest of time

Access to Email

- ▶ Solr search engine already indexes metadata of ingested records and makes records available for download
 - Item must be marked as publishable first
 - Access restrictions set at series level prevent auto-publishing records
 - No staff time to read email one-by-one
 - Conclusion: preserved, but not accessible

Questions?

- ▶ Elizabeth Perkes
 - ▶ Electronic Records Archivist
 - ▶ Utah State Archives
 - ▶ eperkes@utah.gov
- 