Email Archiving at SFU

SFU Archives began testing email transfer in 2012. Over the next several years, the Archives acquired email from a number of SFU accounts. Following agreement with the account holder or their estate, the university's IT Services department (ITS) made backups of the accounts and stored them offline. In April 2018, ITS transferred the backups to Archives by restoring their contents to a dedicated Archives' email account. Archives then exported and converted the email to `mbox` format and ingested the files into the Archives' preservation system (Archivematica) as unprocessed backlog.

In 2019-20, the Archives undertook an arrangement and description project with one of these email accounts (Michael Fellman, fonds F-260). The project established a basic workflow to integrate a number of different software tools for processing and managing email archives. In 2021 and 2022 we were able to acquire email from two additional SFU accounts, and this provided an opportunity to review / revise transfer procedures. Transfer procedures are now in the process of being updated (July 2022) for both the account holder and archivist points of view.

Acquisition thus far has been ad hoc, as opportunity presented, typically at the end of an account holder's SFU career. Some email was transferred by SFU administrators upon retirement or leaving the university; others by the account owner's estate as part of a larger donation of a deceased faculty member's personal archives. To date we have received one transfer of non-SFU email, originating from an organization's various gmail accounts.

The Archives' longer-term goal is to develop a more comprehensive and proactive acquisition strategy. The aim will be to achieve wider coverage – more accounts captured and more frequent transfers of records from them. One-time, end-of-career transfers are better than nothing, but we need processes to support on-going, regular transfers of university email business records, as well as privately donated personal and business correspondence.

The Archives plans to begin processing backlog email account in 2022-23 and, as part of this work, to document our appraisal, arrangement, description and access processes. This site will be updated as work progresses.

Last updated: Jul 22, 2022

Software

SFU Archives uses a number of software programs and utilities in its email archiving process. This page provides a basic list, with links to more detailed documentation (when available) on the Archives Digital Repository Utilities site.

The Archives' desktop computers run on Mac OS and all utilities must be able to run in that environment. Archivematica and AtoM are installed on Linux servers, but the user interface is web-based and OS-neutral.

Microsoft Exchange / Outlook

The current platform for SFU's email system.

- SFU Mail).

ePADD

Open-source software for processing historical email collections. Used by Archives for appraisal, selection, processing, and delivery of access to email archives.

- ePADD project site

Archivematica

Open-source software for digital preservation. Used by Archives to create standardized Archival Information Packages (AIPs) for long-term preservation of transferred email.

- Archivematica project site

AtoM (Access to Memory)

Open-source software for archival description and access, platform for Archives' online catalog of finding aids, SFU AtoM. Used by Archives to provide series-level descriptions of email archives in the context of the creator's fonds. For email, SFU AtoM contains descriptions only; access to the actual messages is provided via ePADD.

- AtoM project site
- SFU AtoM site

OfflineImap

Open-source utility for exporting email from IMAP-based email systems. Used by Archives as the main transfer mechanism, exporting email from a target SFU account; outputs email in `maildir` format.

- OfflineImap project site

ClamAV

Open-source anti-virus utility used by Archives to scan transferred email (`maildir`) for viruses and malware.

- ClamAV project site.
- SFU Archives ClamAV documentation

Emailchemy

Proprietary software able to convert email from / to various formats. Used by Archives to convert OfflineImap output (`maildir`) to `mbox` format for upload to ePADD for processing and Archivematica for preservation.

- Emailchemy software site

Python maildir2mbox script

Customized python script running on command line to convert `maildir` to `mbox`. This was the Archives' original method for getting SFU email in `mbox` format, and it provides an open-source alternative to Emailchemy. Emailchemy with its interface is simpler to use.

Correspondents normalizer utility

Custom FileMaker database to support / automate normalization of an ePADD list of correspondents in an email archive.

Archives Information System (AIS) database

Archives' custom in-house FileMaker database. Used to accession email transfers and register AIPs. Documentation available on the Archives' internal wiki.

Last updated: Jul 22, 2022

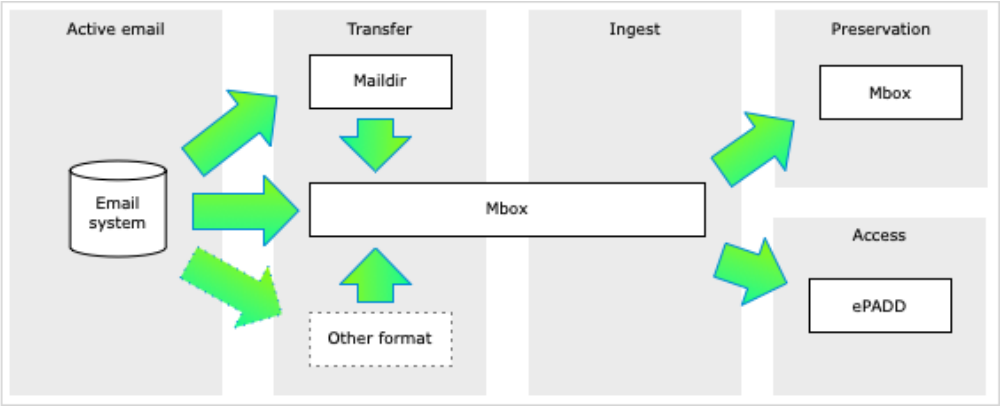
Email Archiving: Formats

Contents

Preferred formats

The Archives' preferred formats for managing email are:

- **Transfer:** `maildir`, `mbox`
- **Ingest:** `mbox`
- **Preservation:** `mbox`
- **Access:** `ePADD`



SFU email platform

The university's email system (SFU Mail) runs on Microsoft Exchange (server) and Outlook (client). SFU switched to Microsoft for email in 2018. From 2009-2018 the university used the Zimbra Collaboration Suite as SFU Connect, and the Archives' first email transfers were from the Zimbra system.

Mbox

`mbox` originated with the Unix operating system as a format for storing email messages, and there are several variants within the `mbox` family. In 2005 the Internet Engineering Task Force (IETF) defined a standard `application/mbox` media type (RFC4155), and `mbox` has become a defacto standard for moving email between different email systems and clients.

A single `mbox` file represents a folder and its contents in an email system.

- It aggregates into a single text file all the messages contained in an email folder, along with their attachments.
- Message headers and body are represented in plain text.
- Attachments are encoded in Base64 as ASCII text appended to the message.

`mbox` files are mainly designed for transfer and storage. While they can be opened with any text editor and the header and body of messages (but not the attachments) are human-readable, access to `mbox` is generally via an email client / reader.

`mbox` is the Archives' preferred **preservation format** due to its wide use, compatibility with multiple email systems, simplicity of structure (as basic ASCII text), and the availability of cross-platform tools that can work with it. But the Archives' digital preservation software, Archivematica, cannot normalize email to `mbox`; this must be done independently prior to ingest. Similarly, ePADD needs email already in `mbox` format for upload. For these reasons, `mbox` is also our preferred **ingest** format.

While some email systems may be able to natively export to `mbox` (e.g. Gmail), others (including SFU Mail) may do so poorly or require the production of an intermediary. SFU Archives uses a tool (OfflineImap) to export email out of the active system in `maildir` format, then converts it to `mbox` with another utility (Emailchemy) or python script.

In general, for a format to be acceptable for transfer to SFU Archives, there must be a tool that allows conversion to `mbox`.

Maildir

Like `mbox`, `maildir` is an email storage format that represents email messages (header, body, attachments) as text files, with attachments encoded as Base64 ASCII. But where `mbox` aggregates all messages from the same mailbox into a single file, `maildir` stores each individual email message (plus attachments) as its own separate text file.

The granularity of `maildir` gives it some advantages. Single messages that have been corrupted or infected with malware can be readily isolated from the others in the same mailbox folder.

But `mbox` is more widely used, and for the Archives `maildir` is an intermediary **transfer format**. Our transfer / export tool (OfflineImap) exports only to `maildir` and the output must be converted separately to `mbox`. Initially, the Archives did not retain the `maildir` version once ingest of the `mbox` was complete. As of July 2022, however, the `maildir` is retained now in backlog pending full processing of the email with ePADD.

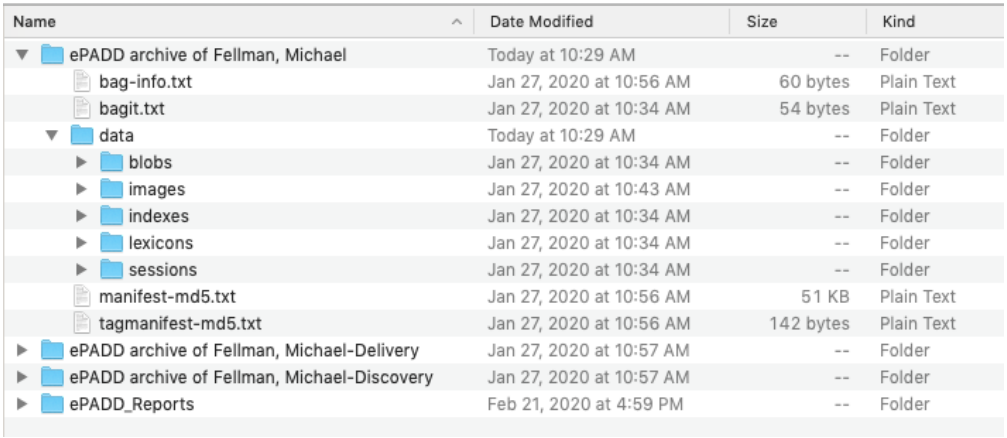
PST files

`pst` (Personal Storage Table) files are a Microsoft-specific format for storing copies of email messages and attachments, but also calendar events, tasks, and contacts. SFU Archives has not acquired `pst` files and it is a non-preferred format.

Our initial analysis is that `pst` files are typically less complete and reliable than `mbox` and that the tools for migrating them to `mbox` are not always robust. The Archives' preference at the present time is to use OfflineImap to export email from Microsoft Exchange / Outlook email systems as `maildir`, then convert the `maildir` to `mbox`. But our experience to date with this method is restricted to working with SFU email.

There may be circumstances where university or privately donated transfers include `pst` files and this is the only format available. When / if this occurs, the Archives will look more closely into tools for dealing with this format. Emailchemy, for example, should be able to convert `pst` files to `mbox`.

ePADD



Name	Date Modified	Size	Kind
ePADD archive of Fellman, Michael	Today at 10:29 AM	--	Folder
bag-info.txt	Jan 27, 2020 at 10:56 AM	60 bytes	Plain Text
bagit.txt	Jan 27, 2020 at 10:34 AM	54 bytes	Plain Text
data	Today at 10:29 AM	--	Folder
blobs	Jan 27, 2020 at 10:34 AM	--	Folder
images	Jan 27, 2020 at 10:43 AM	--	Folder
indexes	Jan 27, 2020 at 10:34 AM	--	Folder
lexicons	Jan 27, 2020 at 10:34 AM	--	Folder
sessions	Jan 27, 2020 at 10:34 AM	--	Folder
manifest-md5.txt	Jan 27, 2020 at 10:56 AM	51 KB	Plain Text
tagmanifest-md5.txt	Jan 27, 2020 at 10:56 AM	142 bytes	Plain Text
ePADD archive of Fellman, Michael-Delivery	Jan 27, 2020 at 10:57 AM	--	Folder
ePADD archive of Fellman, Michael-Discovery	Jan 27, 2020 at 10:57 AM	--	Folder
ePADD_Reports	Feb 21, 2020 at 4:59 PM	--	Folder

ePADD software allows an archivist to work on a collection of `mbox` files to appraise, select, and describe email archives. It provides a number of curation features, including support for message annotation and tagging, grouping by correspondent, entity analysis, and management of restrictions.

Processed materials are output as three separate ePADD packages:

- the processed email;
- the discovery copy: a redacted version displaying only names of entities, all other text being removed; intended to be suitable for online dissemination;
- the delivery copy: the full email for messages cleared for access; intended to be delivered in a stand-alone, offline envirnment.

Each ePADD package is stored as a Bag, following the BagIt specification. Within the Bag `data` folder, the content is stored in an ePADD-specific structure in the sense that it can only be accessed and rendered via ePADD software. For this reason, SFU Archives treats the set of ePADD packages as the **access copy**, but continues to rely on `mbox` as the long-term **preservation copy**. This does mean, however, that the ePADD-specific curation features added during processing are only retained in the access copy and **not** in the preservation `mbox` file(s).

ePADD itself can generate `mbox` files, and ideally the preservation copy would be an `mbox` created by ePADD after selection and processing is complete. However, a significant limitation of the current version of ePADD (v8) is that it creates only a single `mbox` file without preserving the folder structure of the original transfer (this folder structure is preserved in the ePADD version, but is lost on export to `mbox`).

The recent (2022) ePADD+ project should address this issue by supporting the production of preservation copies that retain original folder structure. Until this is implemented, the Archives will continue to rely on the `mbox` files generated at transfer as the preservation copy. This has a significant downside, as it means that the preservation copy includes messages and attachments that were designated for destruction during ePADD appraisal, selection, and processing. If a future release of ePADD is able to retain original folder structure in the preservation `mbox` files, the Archives will use these to replace the pre-processed versions currently ingested to Archivematica.

Attachments

Attachments are written to `mbox` files as Base64 ASCII text and in ePADD retained in their original file formats. In both cases, no preservation actions are taken on the files themselves, making them at risk of becoming inaccessible in the future.

The Archives' current strategy is to use ePADD's ability to export attachments to create a standalone package consisting of all email attachments. This is then ingested to Archivematica, with full preservation and access normalization applied to all files.

One difficulty with this approach is protecting the link between an attachment and its parent message. In ePADD, all attachments are assigned a unique sequential ID number embedded in the file name. If an attachment in its original file format becomes unreadable in the ePADD access copy, it will be possible to identify the file within the package of attachments and retrieve the preservation or access copy from offline storage.

This seems likely to be a cumbersome operation. Moreover, the ePADD-assigned ID numbers will not be present in the preservation `mbox` version. In general, management of attachments is an area requiring more investigation and work.

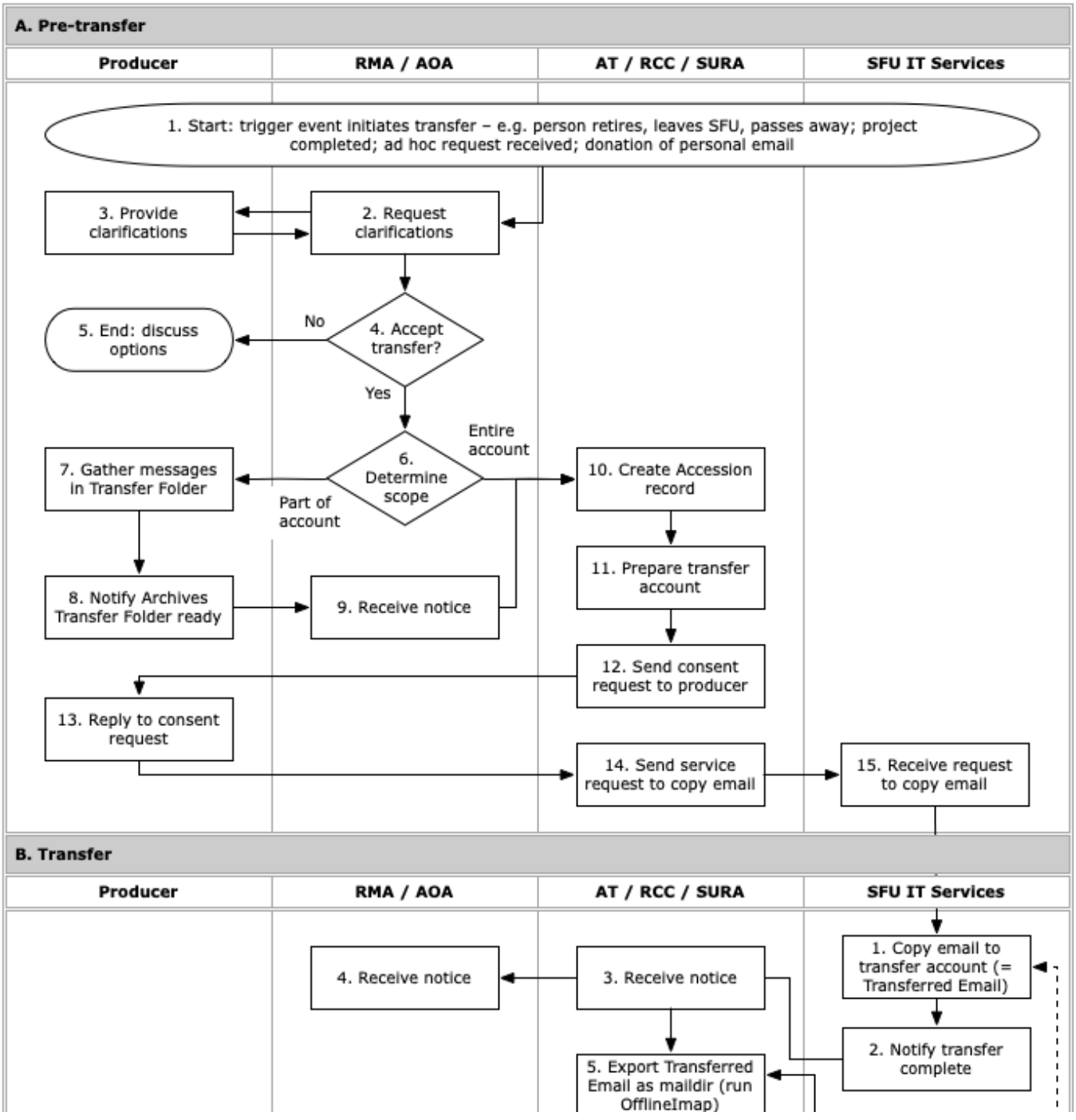
Last updated: Jul 22, 2022

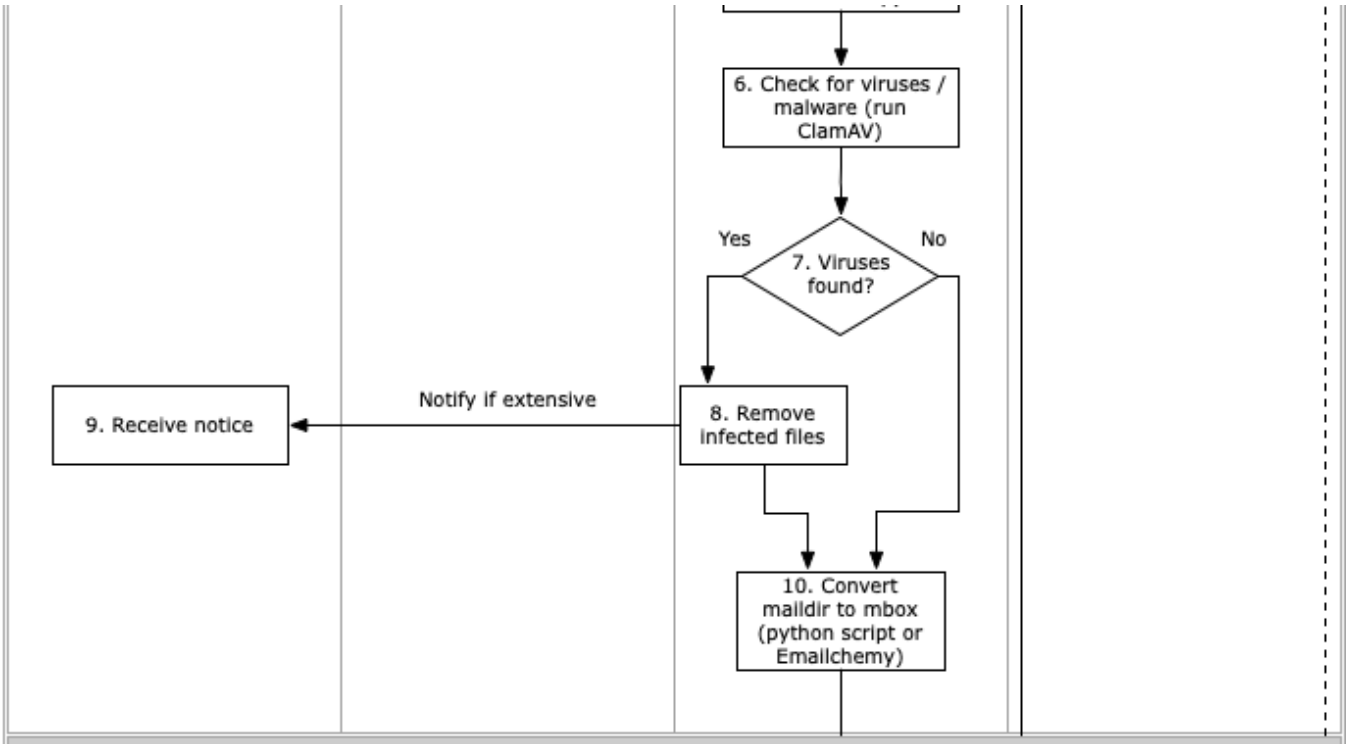
Transfer Workflow

The page outlines in brief the workflow for transfers of SFU email. For more detail (from the archivist's point of view), see the sections on Transfer Workflow.

SFU Archives
Workflow for SFU Email Transfer
Last updated: Jul 18, 2022 (RD)

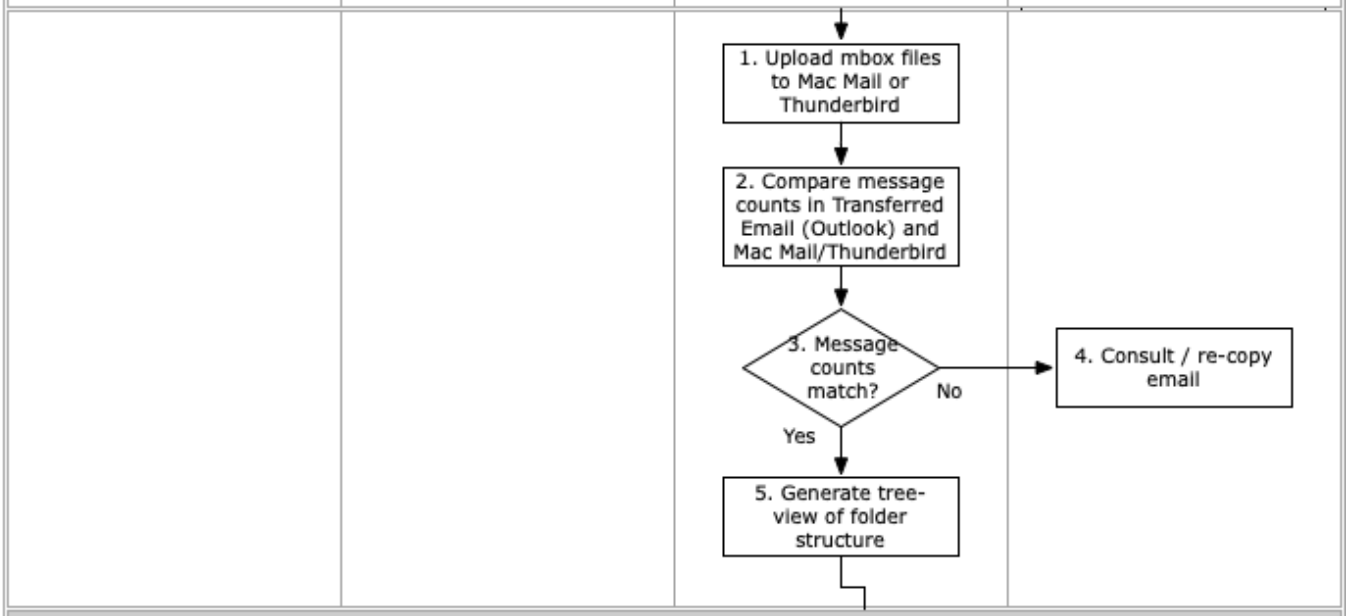
Agents:
Producer = SFU departmental staff, private donors
RMA = Records Management Archivist (for university records)
AOA = Acquisitions and Outreach Archivist (for private records)
AT = Archives Technician
RCC = Records Centre Clerk
SURA = Systems and University Records Archivist
ITS = SFU Information Technology Services





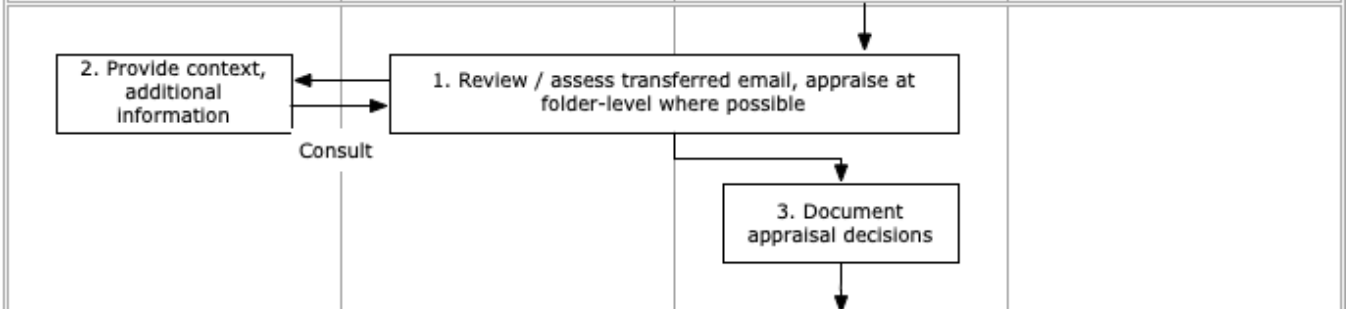
C. Validation

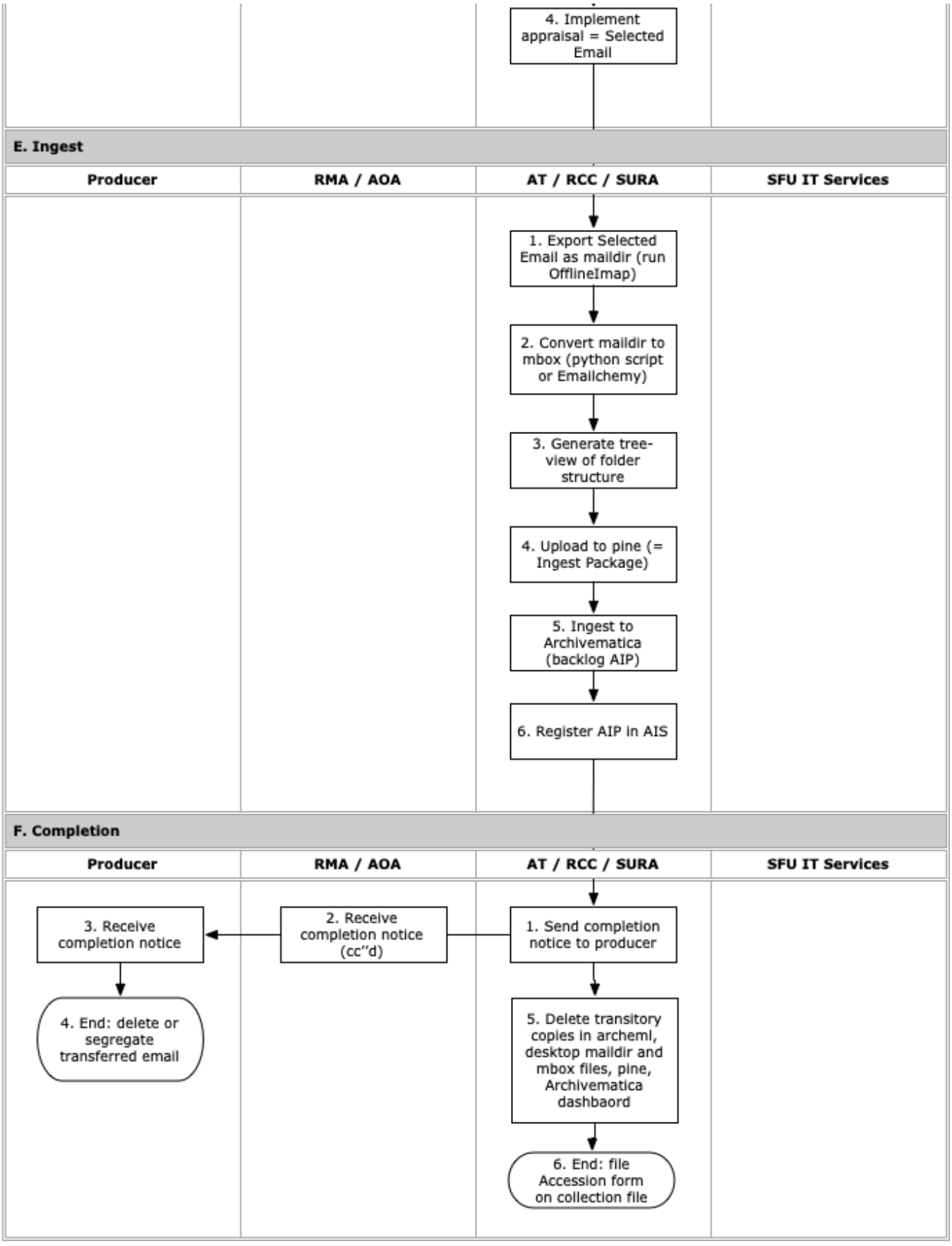
Producer	RMA / AOA	AT / RCC / SURA	SFU IT Services



D. Appraisal

Producer	RMA / AOA	AT / RCC / SURA	SFU IT Services




4. Implement appraisal = Selected Email
E. Ingest
Producer
RMA / AOA
AT / RCC / SURA
SFU IT Services
1. Export Selected Email as maildir (run OfflineImap)
2. Convert maildir to mbox (python script or Emailchemy)
3. Generate tree-view of folder structure
4. Upload to pine (= Ingest Package)
5. Ingest to Archivematica (backlog AIP)
6. Register AIP in AIS
F. Completion
Producer
RMA / AOA
AT / RCC / SURA
SFU IT Services
3. Receive completion notice
2. Receive completion notice (cc''d)
1. Send completion notice to producer
4. End: delete or segregate transferred email
5. Delete transitory copies in archeml, desktop maildir and mbox files, pine, Archivematica dashbaord
6. End: file Accession form on collection file

Pre-transfer

Negotiate a transfer agreement with an account owner ("producer") and determine the scope of transfer (entire account or specific folders). The producer prepares a `Transfer Folder`, the archivist submits a service request ticke to SFU IT Services (ITS).

Transfer

Receive a copy of the targeted email `Transfer Folder` from ITS into the Archives' dedicated email account; use OfflineImap to export the messages + attachments as `maildir`, run a virus scan, then convert the `maildir` to `mbox` format.

Validation

Determine whether transfer / export / conversion was successful (no data lost) by comparing message counts in the transfer account in SFU Mail vs. the `mbox` files when opened in Thunderbird or Mac's Mail email client; generate a tree-view of the folder directory structure.

Appraisal

Where feasible, conduct folder-level appraisal, document appraisal decisions and eliminate folders not selected for long-term preservation.

Ingest

Re-export appraised / selected email, convert to `mbox`, upload to the staging server, ingest to Archivematica and register the AIP in the AIS database.

Completion

Notify the producer that the transfer has been completed, finalize the `Accession record`, and delete all transitory copies from SFU Mail, desktop, and staging servers. The producer may now delete the transferred email if desired.

Last updated: Jul 22, 2022