# EAS Dataflow Text

## Draft: April 5, 2014, W. Gogel and G. Reilly

1. Archivist acquires email

The archivist acquires email through donation or as part of university records. Acquisition methods include receiving portable media such as CDs or hard drives, transferring to EAS from a donor computer, or downloading from remote computers to an archivist’s workstation. [Refer to EAS Best Practices for Archivists document re: how sensitive data is secured at this stage]

1. Archivist creates an EAS “packet”

On their own workstation or the donor’s computer, the archivist arranges email files and folders into a single directory that becomes the EAS packet. [Refer to EAS Best Practices for Archivists document re: how sensitive data is secured at this stage]

1. Archivist transfers the packet to the EAS dropbox

Using a VPN tunnel, the archivist transfers the packet of email files and folders to the secure EAS dropbox file system via secure FTP (sftp). [Refer to EAS Best Practices for Archivists document re: how long to save the original acquired content]

1. Archivist uses EASi to submit the packet

To submit the packet that is waiting in the dropbox, the archivist logs-in to their own account in the EAS administrative interface, EASi. EASi is only accessible using HTTPS via a VPN. In EASi, the archivist can enter metadata to describe the packet and then submit the packet for ingest into EAS.

1. EAS loads the packet

Once the archivist submits a packet, the EAS loader will detect the packet and move it from the secure drop box file system to a secure storage file system location designated for EAS.

1. EAS imports and processes the packet

The EAS importer picks up the packet from the secure storage file system and runs a series of processes that convert the original packet, creating a copy for further processing. Both the original packet and the converted copy are kept in the EAS secure storage file system. The EAS importer process converts email messages to the EML standard format; copies external attachments; extracts embedded attachments; writes metadata supplied by the archivist to an Oracle database; associates rights metadata restrictions of “Rights basis: risk assessment” and “Secure Storage required: unconfirmed;” and stores the metadata plus extracted email message content (headers and bodies) in a Solr index on the secure storage file system. See XXX document for more details.

1. Archivist engages in archival processing in EASi

Once the packet is imported, an archivist can process the content by editing metadata, aggregating the email and attachments into collections, and deleting email and attachments. These activities modify the metadata in the Oracle database, modify the converted copy (through deletion of emails and/or attachments), and modify the content of the Solr index. Here, as they review the content, archivists can modify the “Secure storage required: unconfirmed” designation by either removing it (for content determined by the archivist to be non-sensitive) or by changing it to “Secure storage required: confirmed” (for content determined to be sensitive).

1. Archivist selects email and attachments to permanently archive and pushes it to the DRS

The archivist marks selected email messages and attachments, along with associated metadata, for transfer to the Digital Repository Service (DRS) for long-term preservation by using the automatic “Push to DRS” function in EAS. Once email messages and attachments are marked for transfer to the DRS, they can no longer be processed in EAS. EAS assigns a unique batch ID to the selected messages and attachments in the Oracle database and to the related items in the Solr index. The metadata status of the items is also updated in the database and index to indicate that they are “ready to archive.”

1. EAS prearchiver prepares batches for DRS loader

The EAS prearchiver reads the Oracle database to identify email messages and attachments that are marked for transfer to the DRS and copies the relevant converted files to an archive queue on the EAS secure storage file system. At the same time, it organizes the files into the batch directory structure expected by the DRS loader and generates the associated metadata descriptor files expected by the DRS loader. Since EAS batches may be larger than the size of batches that DRS can accommodate for loading, one EAS batch may be divided into several DRS deposit batches. Copies of these DRS deposit batches are transferred to the secure DRS dropbox file system via sftp. [Grainne confirm that this is via sftp.]

1. DRS loader processes batch deposits

The DRS loader detects a batch (in this case from EAS) in the DRS secure dropbox. The DRS loader validates the batch against known DRS content models, and validates metadata stored in metadata descriptor files against standard and local metadata schemas. The metadata is written to the DRS Oracle database. A process assigns unique Oracle IDs to the content in the DRS Oracle database; updates the metadata descriptor files with the IDs; and replaces the local file names with their corresponding IDs. Persistent identifiers (URNs) are generated for the content and written to the DRS Oracle database and the metadata descriptors.

The metadata descriptors are written to the DRS standard file system. They descriptors associated with the email content contain the following fields from the email message headers:

* Content type (Grainne – confirm the outstanding question here)
* Message ID
* To
* From
* CC
* BCC
* Origination Date
* Subject Line
* Body Part Content ID (identifier of an embedded attachment that has been extracted) (see ZZZZ doc)

These fields are indexed in the Solr index, which resides on standard storage. No other content from email messages is indexed, for example the full-text of email bodies is not indexed so that it cannot be searched once content is deposited in DRS.

If an email message or attachment has a “Secure storage required” designation of either “confirmed” or “unconfirmed,” the file will be stored in the DRS secure storage file system. If there is no such designation, then the file will be stored in the DRS standard storage file system (*see* 6 and 7 above for more details about the secure storage designation).

DRS currently indexes the full text of all PDF and Text documents regardless of their secure storage designation. Since, for security reasons, the full text of any attachments that have been designated as requiring secure storage should not be indexed, this will be corrected before EAS moves to production: In production, no files that are designated as “Secure storage required” will be indexed.

A DRS process scans for viruses and creates a metadata flag called “admin flag” if a virus is found. The admin flag is added to the DRS Oracle database, the Solr index, and the metadata descriptor.

The status of the batch load into the DRS (either success or failure) is sent by email to all email addresses configured for the account.

1. Archivists manage and download DRS content including email messages and attachments through the DRS Web Admin staff user interface

Once the DRS loader process has completed, authorized DRS accountholders will be able to modify most metadata related to their email collections through the DRS Web Admin, but will not be able to add, remove or change a “Secure storage required” designation. In the current architecture, once content is stored in the DRS, it cannot be transferred from standard to secure storage or vice versa. In the DRS Web Admin, authorized DRS accountholders will be able to download email messages and attachments along with their associated metadata descriptors. This functionality has not yet been implemented for email messages and attachments designated as needing secure storage. (*see* 9 above for more details about descriptors). [Refer to EAS Best Practices for Archivists document re: mediated access here]

1. DRS delivery services render content to web browsers

In the current DRS architecture, the only DRS delivery service capable of delivering email messages is the File Delivery Service (FDS): email messages in secure storage will not be delivered to users. Email messages in standard storage will behave according to the DRS metadata access flag setting. Content with an access flag set to "N" will not be delivered to a web browser; content with an access flag set to "R" will be delivered to a web browser after authentication via a Harvard ID and PIN; and content with an access flag set to "P" will be delivered to any web browser.

It is envisioned that a delivery service for email messages will be implemented in the future.

Email attachments can be of any format, and are deposited to the DRS as one of the following content models: Still Image, Text, Document, Audio, or Opaque. Attachments that conform to the Still Image, Text, Document and Audio content models are capable of being rendered by DRS delivery services, specifically: the Image Delivery Service (IDS), the File Delivery Service (FDS), and the Streaming Delivery Service (SDS). Each of these delivery services checks the DRS access flag before delivering content to users and delivers content only from DRS standard storage (attachments in secure storage will not be delivered to users). [Grainne – confirm that FDS will not deliver out of FDS]

Attachments in standard storage will be delivered according to the DRS access flag setting: content with an access flag set to "N" will not be delivered to a web browser. Content with an access flag set to "R" will be delivered to a web browser after authentication via a Harvard ID and PIN. Content with an access flag set to "P" will be delivered to any web browser.

Attachments that are stored in standard DRS storage and conform to the Text or Document content models will be indexed and searchable by the Full Text Search Service (FTS), and may be delivered by FDS, but with the same delivery behavior based on the DRS access flag setting as described above. Attachments in DRS secure storage will not be indexed and therefore are not searchable by FTS.

If an attachment does not conform to the Still Image, Text, Document and Audio content models, it will be deposited to the DRS as an Opaque object. Any content, including Opaque objects, can be downloaded by authorized DRS accountholders using the DRS Web Admin (as described in #11).

1. EAS Postarchiver cleans-up (annotated below, in parenthesis, with the matching data store identifiers from “EAS Data Flow Diagram”)

For each DRS deposit batch, once EAS detects a successful DRS load report, the EAS Postarchiver deletes: 1) the DRS deposit batches from the DRS secure dropbox (D9), 2) the ArchiveQueue copy from the EAS secure storage file system (D8), and 3) the EAS-converted copies of email messages and attachments from the EAS secure storage file system (D5). The EAS Postarchiver removes the metadata supplied by archivists from the EAS Oracle database (D6) and removes the email message content and metadata from the EAS Solr index (D7). The original submitted packet currently remains untouched in the EAS secure storage file system (D4).