DIGITAL REPOSITORY PROJECT

Repository Infrastructure Requirements

Last revised by: RD (draft 6) **Last revised date:** March 21, 2014

Overview

See diagram (p. 4) for visual representation; numbers here correspond to diagram and entries in specifications chart.

SFU departments transfer materials over the network to **Transfer space server** VM (9); transfers governed by retention schedule.

Private (non-SFU) donors transfer materials over the network to **Archives processing machine** (8); transfers governed by donation agreement.

Large transfers (SFU or non-SFU) can be transferred manually via **external hard drive** (7): network testing to determine file size threshold.

Archivematica pipelines (1, 2, 3) run Archivematica software to ingest materials from **Transfer space** (9) or **processing machine** (8); each pipelines runs on its own VM.

- There are three separate Archivematica pipelines: a **test instance** (1) and two **production sites** (2, 3).
- The test site allows testing against new code not yet available in the public release of Archivematica, as well as against new file formats before definitive ingest via production site.
- One of the production sites (2) is dedicated to processing video: due to large file sizes, this VM needs more RAM and disk space for optimum processing; the other (3) is for documents and images.

Archivematica **production pipelines** (2, 3) send AIPs (originals + preservation copies + metadata) to **AIP storage** (12) disk space.

- AIPs produced by Archivematica **test instance** (1) are for temporary use only and should not be sent to **AIP storage** (12); they can be stored on same VM as the pipeline and deleted when no longer needed.

Archivematica **storage service** (4) is a web-application for managing pipelines and storage spaces; it should be independent of any particular pipeline and have its own VM.

Archivematica Elasticsearch indexes AIPs sent to **storage** (12) allowing for search from an Archivematica pipeline; it can run on each pipeline VM or be configured as its own separate VM (10).

- The advantage of separating Elasticsearch is that it would permit searching all AIPs from all pipelines (otherwise each pipeline can only search its own AIPs).

- A separate **Elasticsearch** VM (10) is a long-term goal; it may be needed as the repository grows and especially if additional Archivematica pipelines are added, but it is not a priority in the short-term.

- Note that the **storage service** (4) can also search all AIPs, but it is a separate application to which only the storage administrator (not all processing archivists) has access.

The Archivematica **production pipelines** (2, 3) send DIPs (access copies + minimal descriptive metadata) to the Archives' access system, **Atom** (5, 6), for further description and public dissemination.

- The Atom production site is configured into separate VMs for the **Atom web server** (5) and the **Atom back-end server** (6).

- **Web-server** (5) = public web interface and DIP storage.

- **Back-end server** (6) = MySQL database for descriptions and Elasticsearch.

The **Archivematica test pipeline** (1) sends DIPs to a **test instance of Atom** (11); these are for temporary use only, there is no need in this case for separate front- and back-end servers.

Local redundant backup (13) should include material temporarily stored on the **Transfer storage space** (9) (in processing), **AIP storage** (12), **DIP copies** (5), and the **database of archival descriptions** (6).

- **Transfer storage space** (9) does not need to be included in geo-remote backup.

Specifications

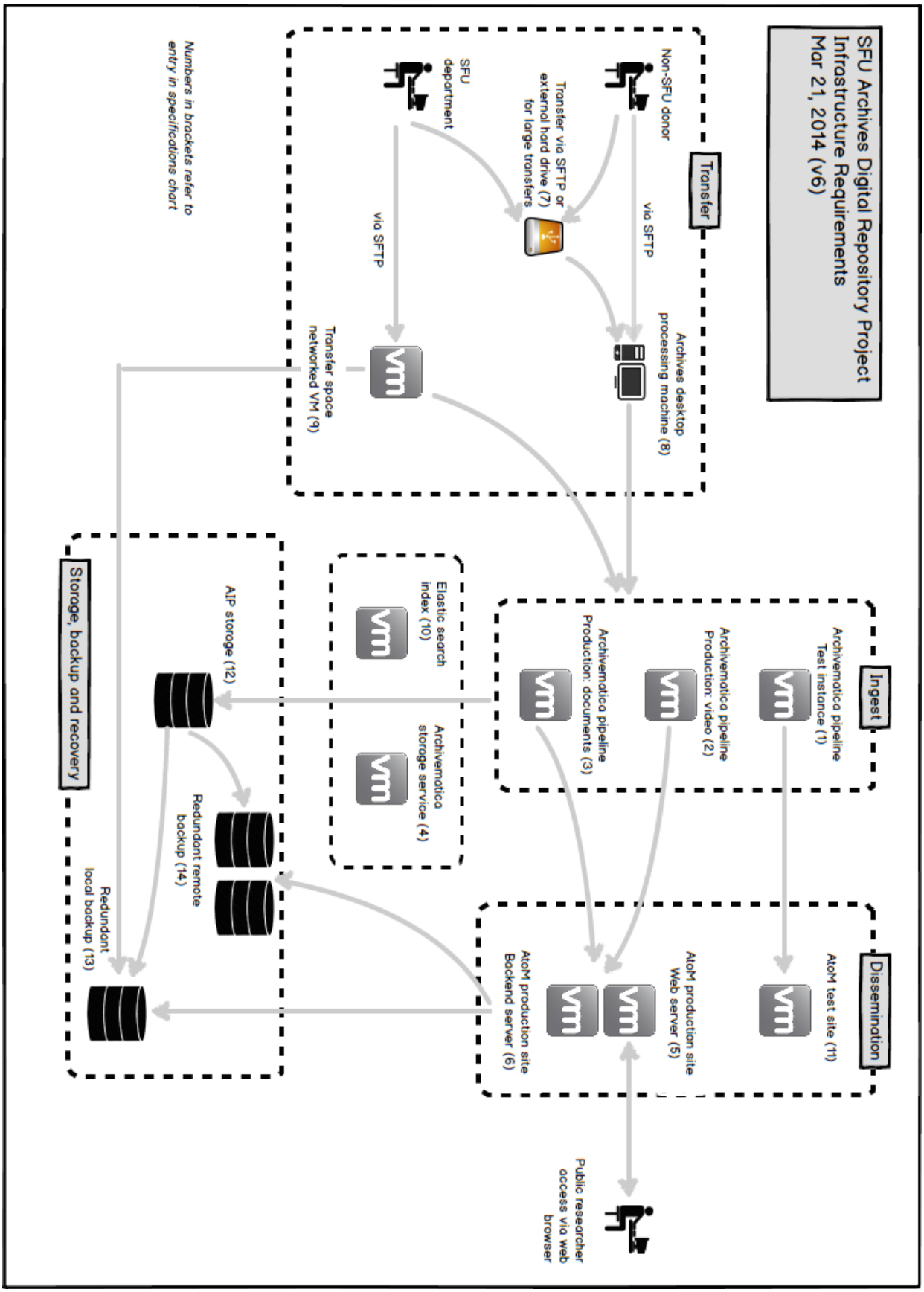
Notes:
- Table is still in-progress, specs for each element not yet complete.
- Elements 1, 5, 6, 8 (shaded) are currently in place.

Element		Type	CPU	RAM	Disk	Purpose
1	Archivematica pipeline 1 – Test instance [earchive.its.sfu.ca]	VM	2	4 GB	1 TB	Test new formats; test new code under development
2	Archivematica pipeline 2 – Production site: video	VM	4	16 GB	4 TB	Definitive ingest of video files
3	Archivematica pipeline 3 – Production site: documents	VM	2	4 GB	600 GB	Definitive ingest of text, sound and image files
4	Archivematica storage service	VM	2	4 GB	200 GB	Manages Archivematica pipelines, storage spaces, and AIP storage

5	AtoM production site – Web server [atom.archives.sfu.ca]	VM	2	4 GB	1 TB	Public web interface and storage of Dissemination Information Packages (DIPs)
6	AtoM production site – Back-end server [atomdb.archives.sfu.ca]	VM	2	8 GB	1 TB	MySQL database of archival descriptions, Elasticsearch index
7	External hard drive	Hard drive			4 TB	Use for larger transfers (>2 GB); use USB3
8	Archives processing machine	Desktop				Use for transfers from external hard drive (7); use as SFTP server for networked transfers from non-SFU donors
9	Transfer space SFTP server	VM			4 TB	Use as SFTP server for networked transfers from SFU departments.
10	Archivematica elastic search index	VM				Indexes AIP storage for searching from Archivematica pipelines (1, 2, 3)
11	AtoM test site	VM				Receive DIPs from test Archivematica pipeline (1); test new code, test imports (xml, text, spreadsheets)
12	AIP storage	Disk	-	-	20-50 TB	Definitive storage of Archival Information Packages (AIPs)
13	Local redundant backup	Disk	-	-	20-50 TB	SFU backup of AIP storage (12), Transfer space (9), AtoM database (6), AtoM DIPs (5)
14	Remote redundant backups	Disk	-	-	20-50 TB each	Redundant copies of AIPs (12), DIPs (5), and archival descriptions (6) to Lower Mainland and geo-remote locations

Diagram

See over.


SFU Archives Digital Repository Project
Infrastructure Requirements
Mar 21, 2014 (v6)
Transfer
Non-SFU donor
via SFTP
Transfer via SFTP or external hard drive (7) for large transfers
Archives desktop processing machine (8)
SFU department
via SFTP
Transfer space networked VM (9)
Numbers in brackets refer to entry in specifications chart
Ingest
Archivematica pipeline Test instance (1)
Archivematica pipeline Production: video (2)
Archivematica pipeline Production: documents (3)
Elastic search index (10)
Archivematica storage service (4)
AIP storage (12)
Storage, backup and recovery
Redundant remote backup (14)
Redundant local backup (13)
Dissemination
AtoM test site (11)
AtoM production site Web server (5)
AtoM production site Backend server (6)
Public researcher access via web browser