LAC Trusted Digital Repository - Virtual Loading Dock

Pam Armstrong, Steve Sekerak, and Marie-Claude Renaud, Library and Archives Canada (Canada)

Abstract

In all domains the amount of digital information is increasing at a rapid rate, this raises crucial questions of preservation. Our intellectual capital, as laid down in educational, scientific, public, cultural and other intellectual resources, is increasingly at risk by the volatile character of digital objects and the rapid developments in information technology.

The growing need for adequate management and preservation of digital information is recognized in many stakeholder communities. The Library and Archives Canada (LAC) is committed to being a Trusted Digital Repository (TDR), to provide reliable and long-term access to the digital documentary heritage of Canada.

Introduction

Increasingly, the documentary heritage of Canada is being born digital and made accessible to Canadians in digital form. The rapid move to a digital environment has changed everything the LAC mandate touches - publishing, government, research, learning, and culture. LAC has therefore set as a primary objective to become a truly digital institution.

The LAC acquires a large scale and broad range of digital content including digital publications, selective websites, large web domains, blogs, electronic government records, digital photos and art, digital audio-visual, geomatics, electronic theses from Canadian universities, digital technical and architectural drawings, private textual electronic records, broadcast data etc. As well, the LAC generates considerable digital content with a large-scale digitization program.

These LAC digital collections are acquired from a variety of domains, through a range of acquisition authorities, with multiple transfer protocols, managed with multiple metadata standards and storage environments and accessed through many different channels.

The Virtual Loading Dock (VLD) is the first step towards the implementation of LAC's TDR. It addresses the OAIS [1] ingest services and is the gateway to the TDR, intended to eventually capture all digital content ingested by LAC. This paper describes the VLD application and the role it plays in assisting the LAC meet its mandate for Digital Preservation. It further describes the business requirements for Legal Deposit and challenges faced by LAC with respect to Digital Preservation. Finally the paper describes the technical design and implementation of the VLD within the LAC TDR.

LAC Mandate for Digital Preservation

LAC is guided by its mandate to preserve the documentary heritage of Canada for the benefit of present and future generations, to be a source of enduring knowledge accessible to all, and to serve as the continuing memory of the Government of Canada and its institutions.

Preservation of the LAC collection of digital materials is based on the broad mandate established by the *Library and Archives of Canada Act* http://laws.justice.gc.ca/en/L-7.7/80647.html. It is governed by the more specific powers outlined in the legislation that relate to the transfer of government and ministerial records of historical or archival value and the transfer of government records at risk, the powers that relate to the Legal Deposit of online publications and the representative sampling of the Internet, and by the provisions in the accompanying *Legal Deposit of Publications Regulations.*

Legal Deposit is applied widely and covers all publications issued in physical format and electronically — effective January 1, 2007, Legal Deposit regulations were extended to include digital publications such as books, newspapers, serial publications and maps.

However, many other types of digital publications are also taken into consideration. Some digital publications require a sophisticated system capacity to ingest and archive and some represent emerging publishing technologies. These types of digital publications will be encompassed by Legal Deposit in the near future.

As the home for Canada's documentary heritage, LAC is alert to collecting and preserving digital publications that may never appear on paper. Through the implementation of its TDR, the LAC has begun to implement the new infrastructure required to guarantee the long-term use of this digital information.

Challenges

The rapidly growing collection of digital and digitized content at LAC needs to be properly managed within a comprehensive set of processes, tools and repositories. Some of the key challenges facing the LAC include;

- **Complexity of Legacy Holdings** — the combined roles of both Library and Archives add to the complexity of the holdings. The digital documentary heritage materials acquired or produced by LAC that form part of the LAC digital collection and must be managed include:
  - Digital publications, either published on physical carriers such as diskettes, compact discs, and CD-ROMs, or published online through the Internet.
  - Digital records, whether received digitally, or on a physical carrier.
  - Web sites, whether collected individually, or as part of a broader harvest of selected domains on the Internet.
  - Digital materials created to enable increased access, these are, digitized copies of traditional format materials contained in the LAC collection.
  - Digital materials, created as a result of the conversion of LAC collection materials in obsolete formats to digital formats, where the new digital version replaces the original version of the material.

- **Capacity Planning** — tremendous amount of digital material of all types is being produced and an exponential pattern of growth is occurring in many areas. As part of its new mandate, LAC began to harvest the web domain of the Federal Government of Canada starting in December 2005. The harvested website data is stored in the Government of Canada Web Archive - http://www.collectionscanada.gc.ca/webarchives/index-e.html. Currently, there are nearly 4 terabytes of data comprising 100 million digital objects accessible to the public. This represents only a small portion of the entire LAC digital collection.

- **Persistent Identification** — ensuring the persistence of information in our digital collection and providing for future accessibility is a complex and expensive undertaking. Implementing a persistent identifier scheme, such as Archival Resource Key (ARK), as well as other international and industry standards will assist LAC in meeting this challenge.

On the path of a Trusted Digital Repository

Recognizing the challenges of digital preservation has propelled the LAC to adopt a new business framework; the TDR. The LAC has committed to a multi-year project to develop a suite of TDR business and technology services to establish a reliable, flexible, integrated digital preservation infrastructure. The LAC TDR is based on the OAIS reference model [1]; it will provide a set of trusted services that provide reliable and persistent access to, along with reliable storage and long-term preservation of, the digital collections at LAC.

The TDR also includes the common set of business functional processes and operations needed to manage the digital objects, the information, people, organization as well as the governance needed to achieve the goals of the TDR. Accordingly, new digital governance bodies have been established within the Institution and new policies developed.

The first step towards the implementation of the LAC TDR is the development of the Virtual Loading Dock (VLD) to enable the capture (ingest) of digital assets. Over the long-term, the intent is for the VLD to capture and ingest all born digital and digitized assets into the LAC collections — whether they are submitted by suppliers, publishers, government departments or donors on physical media, by email, by electronic transfer (FTP, OAI harvester), by web form, or manually collected by LAC staff.

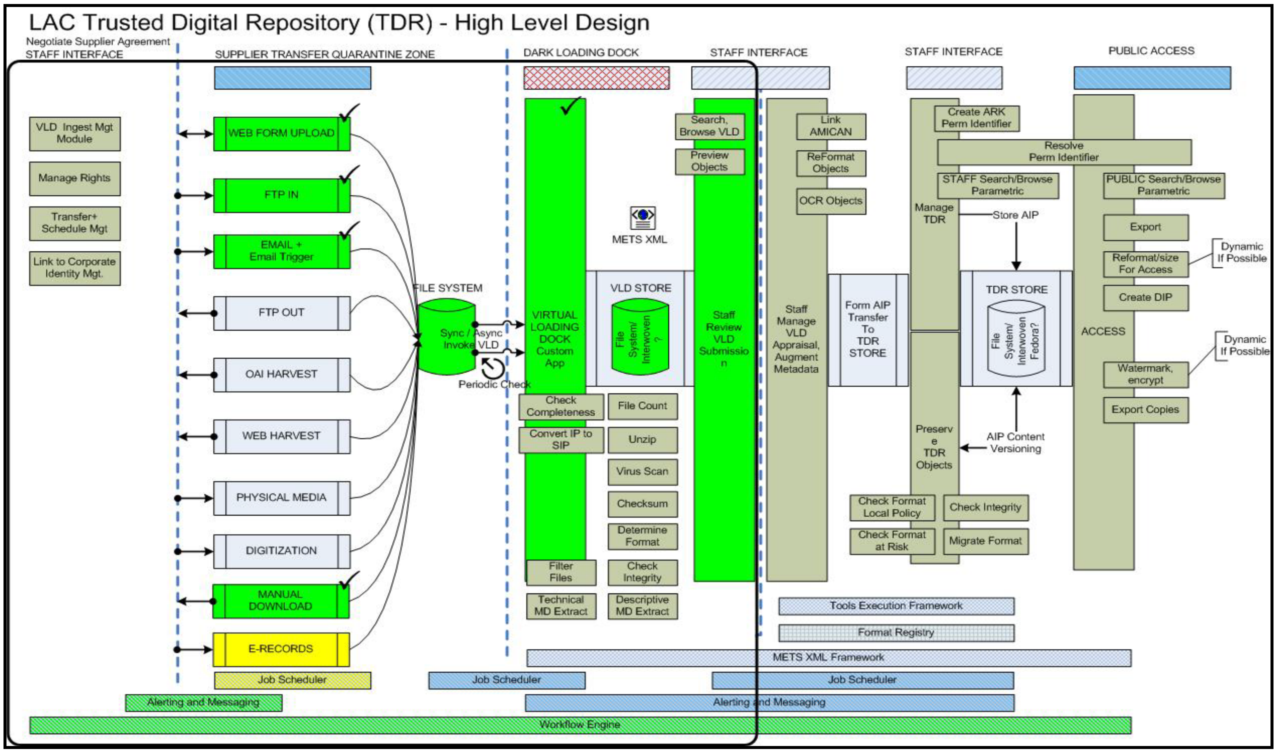

Figure 1. *TDR High-Level Design*

The development of these trusted services closely follows international standards, best practices and guidelines for ensuring the integrity, authenticity and ability to view digital assets within the trusted digital repositories. Following are the metadata standards as well as open protocols and tools currently in use within the implementation of LAC's TDR.

Metadata standards

- METS - Metadata Encoding Transmission Schema [2];
- MODS - Metadata Object Description Schema [3];
- MARC - Machine Readable Cataloguing [4];
- PREMIS - Preservation Metadata: Implementation Strategies [5];
- METS Simple Rights Schema [6];
- Dublin Core [7]; and
- The Government of Canada Records Management Metadata Standard [8].

Open protocols and tools

- OAI – Open Archive Initiative metadata harvester [9];
- PureFTP – Supports LAC security zone [10];
- HTTP Apache, Tomcat [11];
- PHP Hypertext processor[12];
- SOAP [13] + REST [14]- Web services;
- JHOVE – Metadata extract + checksum creation [15];
- DROID – Digital object recognition [16];
- Pronom – File format registry [17];
- Heritrix– Web harvesting crawler [18];
- Wayback Machine – Web archive access viewer [19];
- LDAP – Authentication and role management [20]; and
- Search protocols – Z39.50 [21], OpenURL [22].

The high-level design of LAC's TDR is depicted in Figure 1. The highlighted area indicates the scope of the VLD.

The Virtual Loading Dock

The OAIS Submission Information Package (SIP) processing is the overall process flow for this iteration of the VLD; addressing LAC's requirements for legal deposit and digital published heritage. The VLD is designed to receive digital assets, validate the integrity of the assets, extract technical and descriptive metadata about the assets and prepare the SIP. A SIP is comprised of one or more digital object files and the metadata describing those files within a standards-based representation. Assets are stored in the VLD until they are appraised to determine if they should be part of LAC's digital collection or should be discarded.

The solution is built as custom components and configurations developed using an underlying base of commercial off-the-shelf software products and leveraging open-source technologies. The solution manages the lifecycle of digital assets in three distinct phases:

- **Pre-Ingest** - negotiation of publisher, supplier and publication information which shall be included within the TDR.
- **Ingest** - intake of digital assets through multiple ingestion connectors.
- **VLD Store** - adherence to metadata requirements, validation and processing of assets ingested; the VLD Store serves as the repository of these assets.

The phases are mapped to high-level functional processes within the solution as shown in the VLD Functional Component Architecture diagram in Figure 2. The current iteration of the VLD solution is addressing the ingestion workflow for digital publications.

VLD Functional Components

The following are brief descriptions of the various VLD functional components depicted in Figure 2.

1 Ingest Manager

- Acts as a gatekeeper for moving assets into the VLD. The main purpose of the Ingest Manager is to consolidate assets received from all ingestion connectors into a central quarantine zone, and perform minimal quality assurance before scheduling asset transfer to the VLD.

2 SIP Processing Module

- The SIP module provide the following functionality:
    - It allows for the ingestion of an asset from a Publisher. It secures and verifies the asset;
    - It decompresses/disaggregates the asset into its parts;
    - It validates the asset parts;
    - It generates and pulls together all of the metadata for the asset;
    - It generates the METS;
    - It stores the asset and its associated metadata in the store.

3 Ingestion Connectors

- The ingestion connectors receive publisher produced assets in a connection specific manner. They include www, FTP-IN, email, and Manual (Physical Media and LAC Retrieval).

4, 5 METS Transformation Module (Metadata Handling)

- The METS Transformation module is a component to consolidate the technical, descriptive and administrative metadata extracted from assets during SIP processing and create a valid METS record conforming to the LAC METS specification.

4, 5 METS Update Module (Metadata Handling)

- To provide a mechanism whereby a SIP may be reviewed by LAC staff for the purposes of augmenting or modifying metadata associated with an asset.

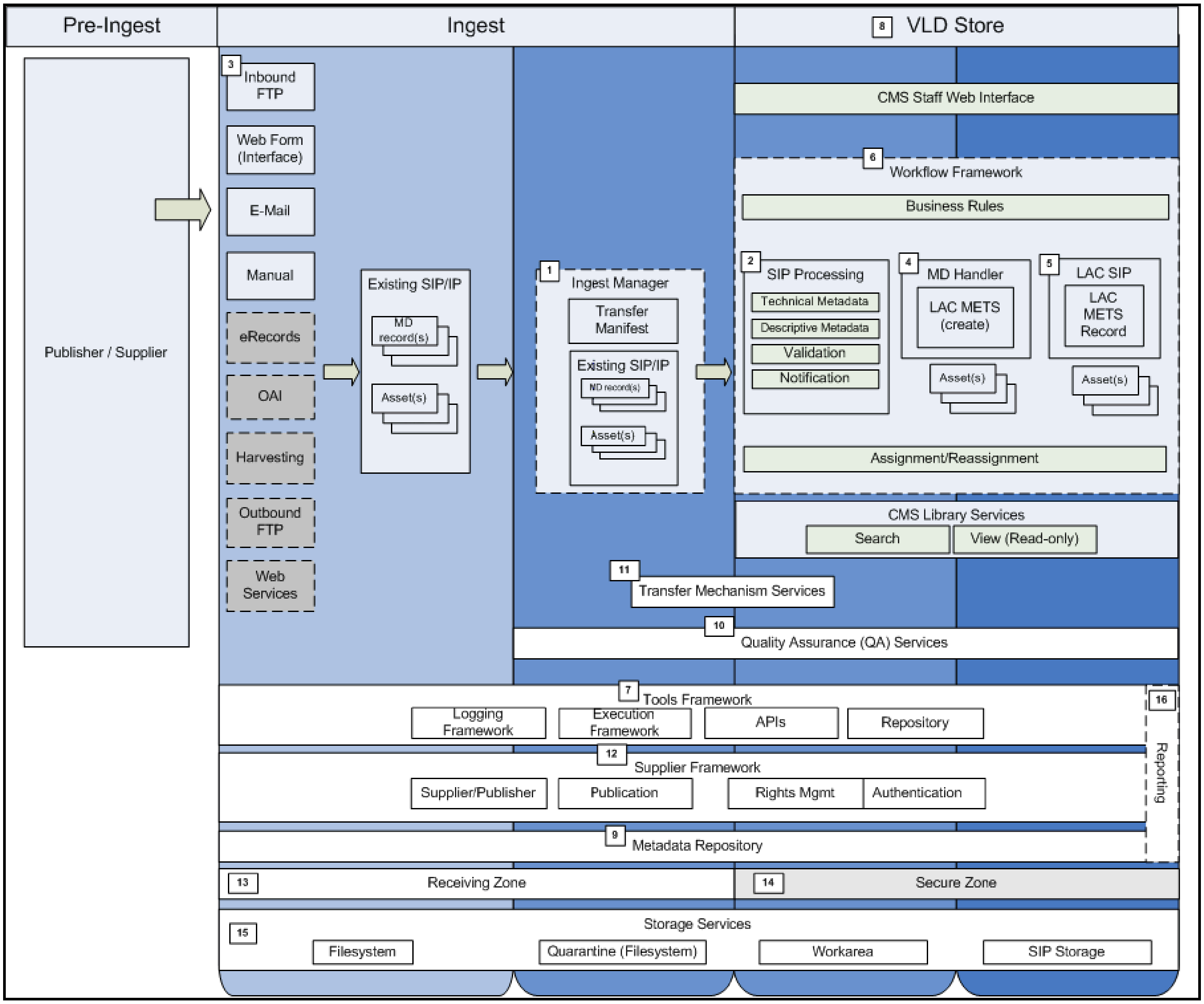

***Figure 2.** Functional Component Architecture*

6 Workflow Framework

- Controls the workflows which will be used throughout the VLD. The design goal of this module will be to control the main processing flow of an asset from ingestion to final SIP storage.

7 Tools Framework

- Provides a repository for, and APIs to query, information concerning the various tools available to perform specific tasks within the VLD. This framework allows for tools to be added and replaced within the tools execution framework with a minimum of recoding effort. VLD tools should, wherever possible, be open-source, 3rd party tools with no custom coding. The tools framework is comprised of the following components:
- Repository - a set of tables to define the tools, tool steps, functions, file formats and a mapping between these entities
- SOAP and PL/SQL APIs to query the tools framework and retrieve the information required to perform a function
- Execution Framework - a series of modules (PERL) to control the execution of wrapper scripts
- Logging – all tools will have their reportable events and output captured and logged from PREMIS.

The following tools are currently in use:

- MD5 - performs an checksum against the ingested assets for verification that asset manipulation (file transfers) does not alter the asset or lose data
- McAffee- performs a virus scan of all ingested assets
- JHOVE - physical asset format and integrity validation

- Interwoven MetaTagger 4.1.0 - Descriptive metadata extraction
- DROID - automated batch identification of file formats - accesses the PRENOM DROID registry hosted as a service by the UK Archives
- WinZip - decompression of packaged assets
- Interwoven Open Deploy - asset transport between the VLD security zones

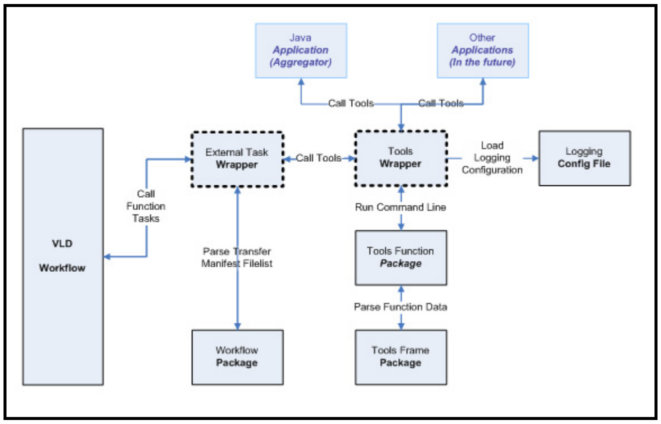

Figure 3. *Tools Execution Framework*

8 Virtual Loading Dock (VLD) Store Module

- The VLD Store Module handles the underlying storage structure to support the processing of a SIP.

9 Metadata Repository

- Metadata extracted from assets as they move throughout SIP processing within the VLD will be stored in a number of tables in the Oracle database. These data elements will be a combination of normalized records, XML fragments and complete XML documents. Currently the metadata will be stored in the manner most effective for extraction and repurposing of that information for the creation of the METS record.

10 Quality Assurance (QA) Services

- A term which refers the collection of functions required to identify, verify, validate and extract metadata from submitted assets. The functions will query the Tools Framework to establish which tool is required to perform that function against a specific asset or a collection of assets.

11 Transfer Mechanism Services

- Controls the transfer of assets from the Ingest storage area to the VLD store. This transfer must be initiated from within the Secure Zone in order to adhere to LAC security standards.

12 Supplier Framework

- External system services which includes a collection of components that define the “who”, “what”, “where”, “when” and “how” of a Supplier and their associated assets.

13 Receiving Zone

- The receiving zone is a network segment which is distinct and separate from the LAC secure zone where protected external facing services reside. The receiving zone has been created for the ingestion as a proactive measure for receiving assets. Every asset ingested into the VLD will be quarantined and virus scanned. Virus scanning must be disabled on the systems within the receiving zone to comply with the Trusted Digital Repository mandate. Virus scanning will be performed by the Ingest Manager module using the scanning tools defined in the Tools Framework. Every file is quarantined, and only those files that pass ALL checks performed by the VLD are then processed by the Ingest Manager to the secure zone, and stored in the VLD Store. For the purposes of the VLD, the receiving zone will house the external ingest servers which include the FTP-IN, email and WWW servers. The receiving zone is protected from the Internet by a firewall and is further segmented from the secure zone by a second firewall. Authentication via the LAC’s corporate identity management service is required from within the receiving zone. Refer to Figure 5 VLD CIM Integration.

14 Secure Zone

- The LAC secure zone is a network segment where protected LAC servers reside. The secure zone contains servers required for the provisioning of internal LAC services and, for the purposes of the VLD, will house the components required to support the Workflow Framework.

15 Storage Services

- Assets are ingested into the VLD from different ingestion connectors (FTP-IN, Web Form, FTP and Manual). The storage services allow these assets to move from the Receiving Zone (File System), to a Secure Zone (VLD Store).

16 Reporting Services

- Provides a mechanism whereby specific pre-defined metrics concerning the processing of assets through the VLD can be generated.

Integration of the VLD Functional Components

The following sections and figures further explain the integration between the different functional components of the VLD solution.

Ingest Manager, Ingestion Connectors and the Tools Framework.

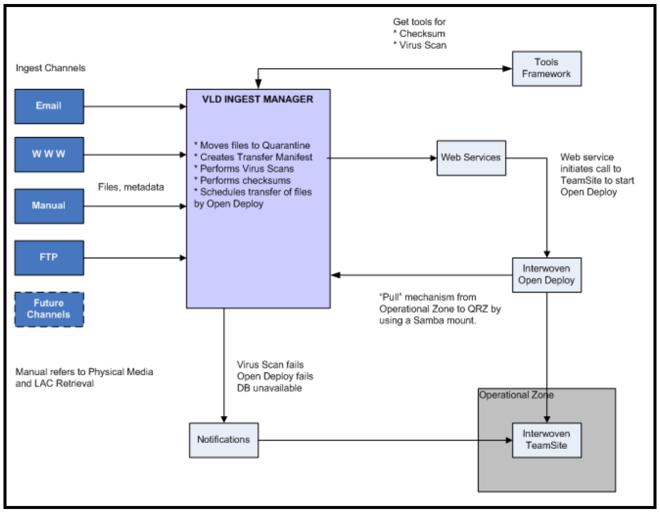

Figure 4. *Ingestion Connectors*

Metadata / METS Handling

The diagram in Figure 5, illustrates the sources and flows of metadata through the system. It explains the original sources of metadata through the online forms as well as system-extracted metadata. It also serves to demonstrate the data flow that results when an LAC Internet Unit staff member edits metadata for a specific asset, and the resulting update in METS.

Metadata regarding a publication asset is derived from the following sources; Publisher Profile, Publication Profile and Extracted Metadata.

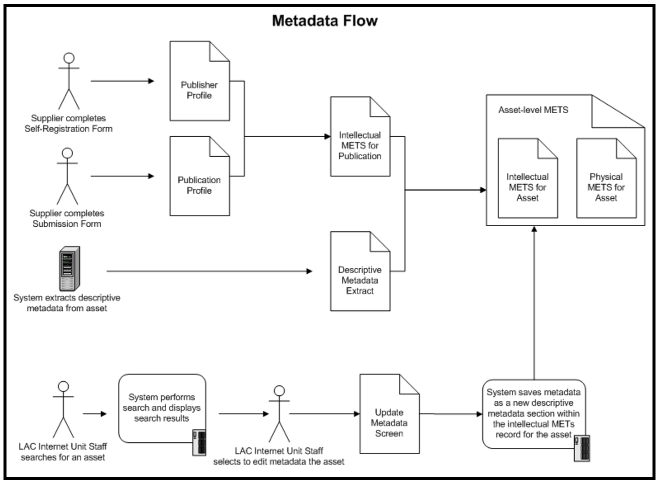
Metadata Flow
Supplier completes Self-Registration Form
Publisher Profile
Supplier completes Submission Form
Publication Profile
Intellectual METS for Publication
Asset-level METS
Intellectual METS for Asset
Physical METS for Asset
System extracts descriptive metadata from asset
Descriptive Metadata Extract
LAC Internet Unit Staff searches for an asset
System performs search and displays search results
LAC Internet Unit Staff selects to edit metadata the asset
Update Metadata Screen
System saves metadata as a new descriptive metadata section within the intellectual METs record for the asset

Figure 5. *Metadata / METS Handling*

Extracted Metadata is processed to extract applicable XMP or MIX data for METS Descriptive sections. Other specific elements will be driven by the Publication Publisher Profile. The purpose of the METS transformation module is to collect the various metadata elements from the discrete repositories and map that metadata into the LAC implementation of the METS schema.

The LAC METS implementation tracks the metadata and events surrounding intellectual entities (a book or serial publication) and the physical files that comprise those intellectual entities. Publication profiles and Publisher profiles will also feed this METS profile.

The generation of a METS record takes place after automated metadata extraction and (before/after) manual validation of an asset.

The modules required to consolidate and assemble the METS record are being coded in PERL to make use of its native file handling, excellent XML processing capabilities and small execution footprint. Since every asset that flows through VLD will require at least 1 METS file (for physical attributes) and potentially 2 (for intellectual attributes), the recommended approach does not included Java processing in order to remove the overhead of initiating the JVM.

Contact Information Management Integration

LAC has a centralized contact information management system (CIM) to provision for integrated authentication and user metadata storage across multiple systems. The VLD provides the ability to authenticate suppliers and access specific metadata elements about the supplier through the various ingestion connectors.

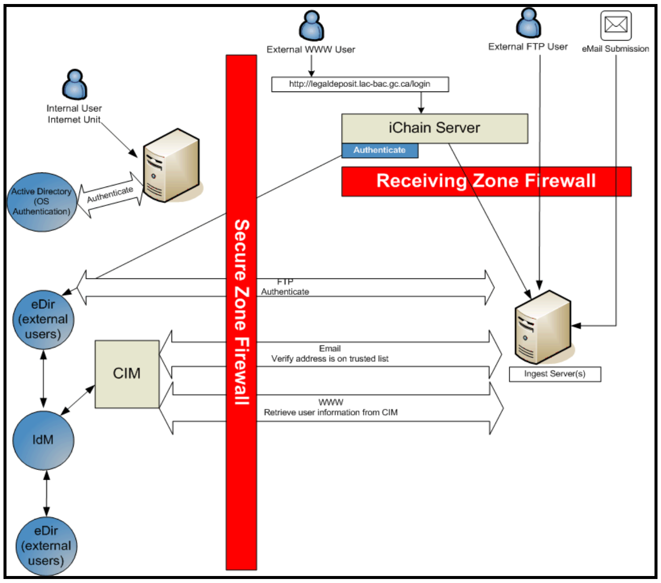

Figure 6. *CIM Integration*

Conclusion

The following are key activities LAC will be addressing over the next year:

- Defining the architecture of it's TDR (post-SIP)
- Implementing the first iteration of a complete TDR for digital publications
- Implementing the ingest service for Government Archives
- Investigating the possibility to use the GCWA service as a means for Government Legal Deposit
- Integrating mass digitization efforts to the VLD/TDR

Acknowledgements

Special thanks to Pam Armstrong, Manager, Digital Repository Services and Standards Office, and Steve Sekerak, Enterprise Architect; LAC's TDR development and implementation would not be possible without their strong leadership.

The Institution would also like to thank the team at Deloitte Touche for their effort and contribution during the development and implementation of the VLD.

References

[1] OAIS, Reference Model for an Open Archival Information System (OAIS). CCSDS 650.0-B-1, Blue Book, January 2002 http://public.ccsds.org/publications/archive/650x0b1.pdf

[2] METS (Metadata Encoding Transmission Schema) http://www.loc.gov/standards/mets

[3] MODS (Metadata Object Description Schema) http://www.loc.gov/standards/mods

[4] MARC (Machine Readable Cataloguing) http://www.loc.gov/marc

[5] PREMIS (Preservation Metadata: Implementation Strategies), http://www.loc.gov/standards/premis

[6] METS Simple Rights Schema http://www.loc.gov/standards/mets/news080503.html

[7] Dublin Core http://dublincore.org

[8] Government of Canada Records Management Metadata Standard, http://www.lac-bac.gc.ca/information-management/002/007002-5002.27-e.html

[9] OAI (Open Archives Initiative), http://www.openarchives.org

[10] PureFTP, http://www.pureftpd.org/project/pure-ftpd

[11] HTTP Apache, Tomcat, www.apache.org

[12] PHP Hypertext Processor http://www.php.net/

[13] SOAP (Simple Object Access Protocol), http://www.w3.org/TR/soap/

[14] REST Architecture, http://rest.blueoxen.net/cgi-bin/wiki.pl

[15] JHOVE, (JSTOR/Harvard Object Validation Environment) http://hul.harvard.edu/jhove/

[16] DROID, (Digital Record Object Identification) http://droid.sourceforge.net/wiki/index.php/Introduction

[17] PRONOM Online Registry http://www.nationalarchives.gov.uk/pronom

[18] Heritrix, (Open Source Web Crawler) http://crawler.archive.org/

[19] Wayback, (Open Source Web Archive Access) http://archive-access.sourceforge.net/projects/wayback/

[20] LDAP Lightweight Directory Access Protocol) http://ca.php.net/ldap

[21] International Standard ANSI/NISO Z39.50, http://www.loc.gov/z3950/agency/

[22] International Standard ANSI/NISO Z39.88 OpenURL, http://www.niso.org/kst/reports/standards?step=2&gid=None&project_key=d5320409c5160be4697dc046613f71b9a773cd9e