

Going with the Flow

Digital Archiving Workflows

ACA 2018

Presenters

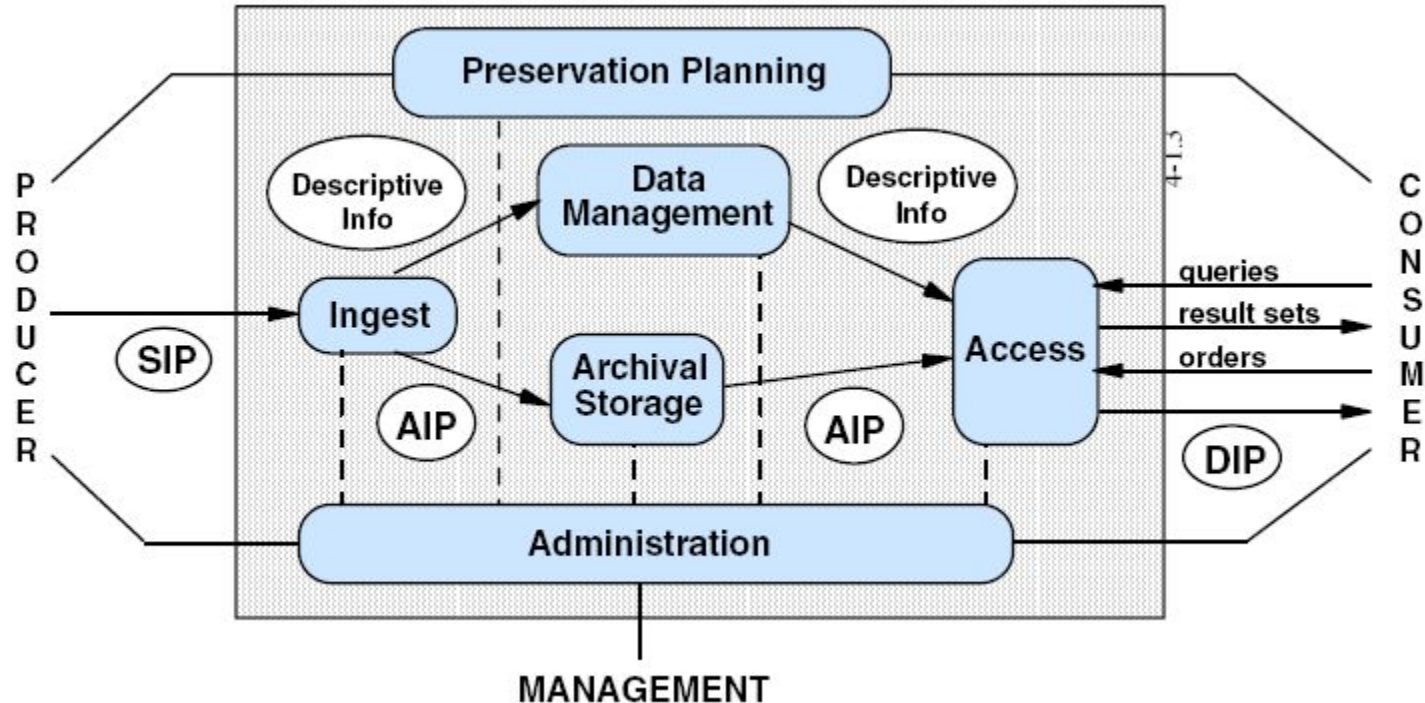
Jess Whyte	Digital Preservation Intake Coordinator Librarian, University of Toronto
Caylin Smith	Legal Deposit Libraries Senior Project Manager, British Library and Legal Deposit Libraries Committee
Krista Jamieson	Digital Archivist, University of Alberta
Grant Hurley	Digital Preservation Librarian, Scholars Portal
Bridget Whittle	Digital Archives Librarian, McMaster University

Outline

- Open Archival Information System (OAIS)
 - Pre, During, and Post
- Digital Archiving Needs
- Institutional Context & Scope
- Acquisition
- Metadata
- Submission Information Package (SIP) and Ingest
- Preservation
- Archival Information Package (AIP) Storage
- Dissemination Information Package (DIP) and Access

OAIS

Open Archival Information System (OAIS) model



Pre-OAIS

Acquisition

Metadata

- Accession
- Descriptive
- Rights Management

OAIS Steps:

Package SIP

Ingest

Preservation & Processing

AIP Storage

DIP Creation & Storage

Post-OAIS

Access

- Rights management
- Copyright
- Privacy & freedom of information, etc.
- Appropriate access: content vs audience

Discovery

Digital Archiving Needs

- Content Acquisition:
 - Donations of born-digital materials
 - Digitization of analog materials
- Storage
- Preservation
 - Bit level (are the 1s and 0s still there?)
 - Renderability (can you read the file? Does it look right?)
- Discovery & Access
 - Rights management
 - Copyright
 - Privacy & freedom of information, etc.

Institutional Context & Scope

University of Toronto - Context



Digital Preservation Unit (2 FTE + 3 project staff)

Digital Archivist (UTARMS, 1)

Archivists and Collection Holders from across UTL (lots)



Special Collections and Archives: born-digital materials and digitized content (not electronic records)

General Collections: licensed/published digital content (not already under consortial control)



Preservation Platform: Tivoli Storage Management (TSM) system with local and offsite storage, various contextual platforms

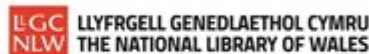
Intake Tools: Primarily linux-based (e.g. fiwalk, diff, rsync, local scripts). Tracking with Jira (for now).

UK Non-Print Legal Deposit - Context (slide 1 / 3)

- Legal Deposit has existed in the UK in some form since 1662
- The Legal Deposit Libraries Act 2003
 - 1 print copy of a publication sent to the British Library and any of the other libraries that claim it
- The Legal Deposit Libraries (Non-Print Works) Regulations 2013
 - Allowed the libraries to collect publications in digital formats
 - Publishers must provide a digital file that is suitable for long-term preservation



Bodleian Libraries
UNIVERSITY OF OXFORD



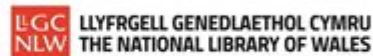
Coláiste na Tríonóide, Baile Átha Cliath
Trinity College Dublin
Coláiste Átha Cliath | The University of Dublin

UK Non-Print Legal Deposit - Context (slide 2 / 3)

- To ensure a national collection of non-print publications;
- To enable an efficient system in which material is archived and preserved in the legal deposit libraries;
- To govern how the deposited copies may be used, balancing the needs of libraries and researchers with the interests of publishers and right holders;
- To facilitate long-term preservation, so that the material may continue to be accessed in future; and
- To ensure long-term viability by requiring both legal deposit libraries and publishers to share the responsibility for archiving without imposing an unreasonable burden on any institution



Bodleian Libraries
UNIVERSITY OF OXFORD



Coláiste na Tríonóide, Baile Átha Cliath
Trinity College Dublin
Coláiste Átha Cliath | The University of Dublin

UK Non-Print Legal Deposit - Context (slide 3 / 3)



- Content and Metadata Processing
- Collection Metadata
- Web Archiving
- Curatorial
- Digital Preservation
- Collection Development
- Dev and test
- Application Support
- Discovery and Access
- amongst others!



Born-digital collections:

- eBooks (PDF and/or EPUB)
- eJournals (PDF)
- UK Legal Deposit Web Archive (WARC)
- Notated music (PDF)
- Geospatial datasets (many formats)



Preservation Platform:

Digital Library System (DLS), a bespoke system built by the British Library

Intake Tools: Publisher Portal, FTP site, Heritrix 3 and Document Harvester (web archive)

University of Alberta - Context



**Digital Archives
Strategy WG** (Digital
Archivist, Metadata, RM,
Digital Preservation,
SysAdmin, Technology
Librarian & AUL)

**Digital Initiatives
(DAMS) & Archives**



Archives: born-digital
materials and digitized
content. Any format &
media type. Institutional &
Private records.



**Collaboration &
Re-use:**
OpenStack Swift storage
'Jupyter' repository
'Pushmi-Pullyu'
Lightweight AIP
Archivematica (TBD)

Access: AtoM (Archives
& Special Collections)

Scholars Portal - Context



Service Provider for members of the Ontario Council of University Libraries (OCUL)

Develops and maintains content-based and member support services

Digital Preservation Librarian + systems & client support staff



Preserves digital licensed materials on behalf of members (journals, books, data)

Develops and sustains **digital preservation services and infrastructure** for the preservation of member-created and managed assets



[TDR](#) for journal content runs on MarkLogic + custom preservation processing pipeline

[Ontario Library Research Cloud](#) (OLRC) for member storage needs

[Permafrost](#) hosted preservation service for processing

Scholars Portal - Permafrost



- Divides responsibility for implementation of OAIS
 - “OAIS Archive” is the subscriber
 - Functional implementation of *aspects* of OAIS: service provider

Scholars Portal:

- Installs and maintains tools, systems and hardware;
- Provides training, consultation, technical support and other resources and guides;
- Provides fixity/reporting data;
- Performs preservation actions over time;
- Implements access requests and assists with integrations

Subscriber:

- Determines content to be preserved,
- Sets policies and procedures,
- Resources the use of the service,
- Validates outputs,
- Dictates preservation actions over time
- Manages access to AIPs & DIPs

McMaster University - Context



Digital Archivist

(Archives & Research Collections, 1)

Support Librarians

(Repository, Technology, 2 [with lots of other commitments])

Other Archivists (3)



Archives & Research Collections

Born-digital materials.
Any format and media type.

Mostly personal records, but some institutional material.



Intake/Validity/Ingest.:

BitCurator (and others), Archivematica

Preservation:

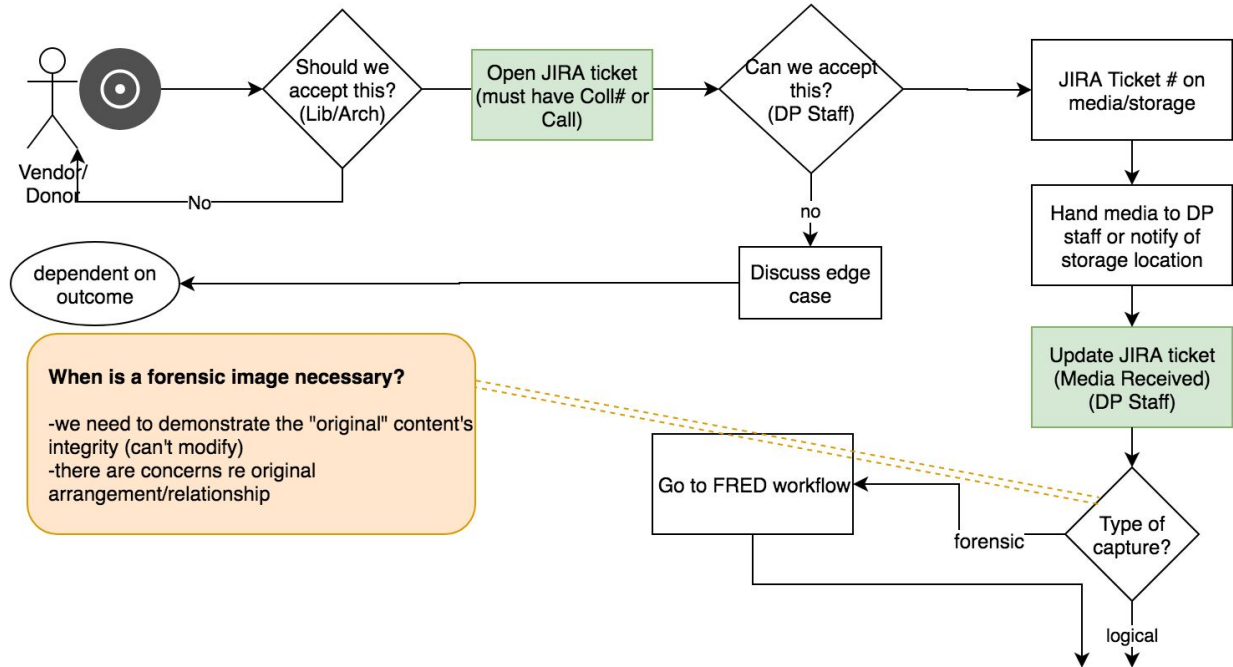
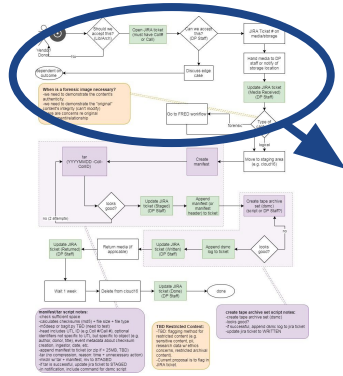
Islandora

Access:

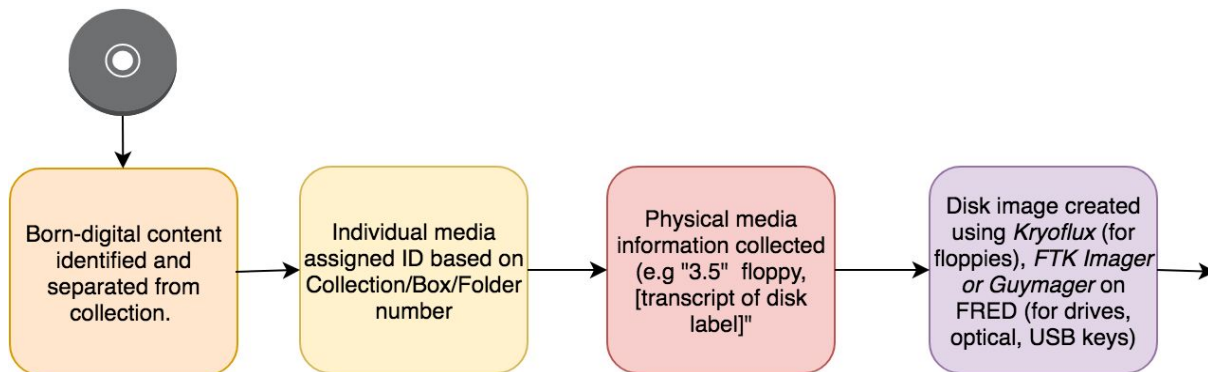
Some combination of AtoM, Islandora, ePadd, others

Acquisition

University of Toronto - Acquisition (General)



University of Toronto - Acquisition (Media)

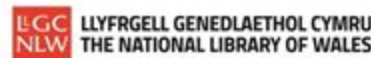


University of Toronto - Acquisition (Reality)



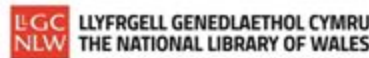
UK NPLD Legal Deposit - Acquisition

- If content is published in the UK, it's in-scope to acquire
 - File format(s) determine if a publisher can be onboarded
- Phase 1 of NPLD addressed short tail of publishers (e.g. large trade publishers, aggregators) as well as long tail of publishers and other content creators
- Phase 2 addressed specialised content (e.g. geospatial datasets and notated music);
- Phase 3 is addressing complex formats not current collected (e.g eBooks created as mobile apps)
- Publishers prioritisation determined by Collection Development and Acquisitions Subgroup (eBooks and eJournals), Legal Deposit Libraries Web Archiving Subgroup, and curators (for maps and music)



UK NPLD Legal Deposit - Acquisition

- For eBooks, eJournals, and notated music, content is deposited directly by publishers or aggregators onto FTP site
- Small eBook publishers (< 1,000 publications per year) deposit using Publisher Portal that was built on SharePoint
- Geospatial datasets are sent to a third party supplier that 'bags' the content using the BagIt standard
- Heritrix 3, an open source, scalable crawler developed by the Internet Archive, is used for yearly UK domain crawl

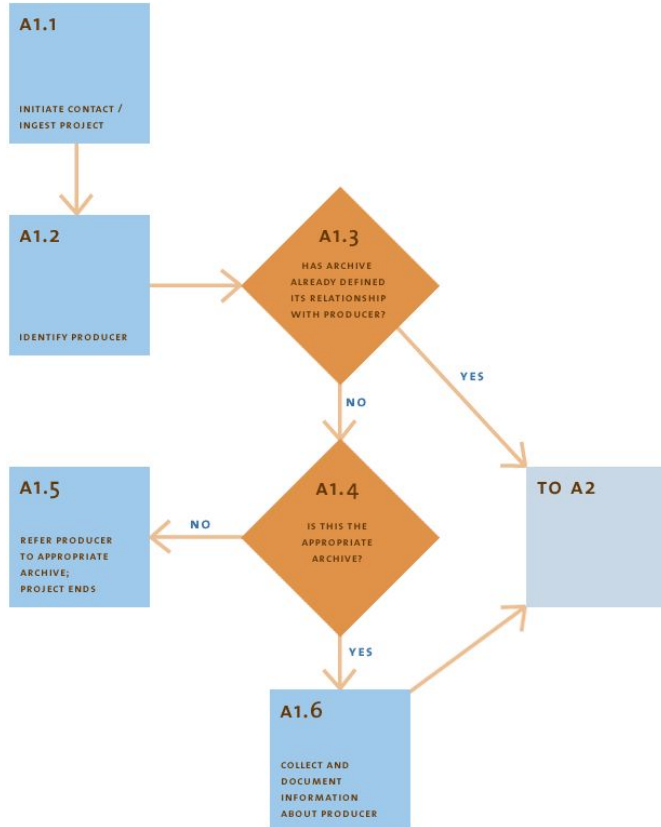


University of Alberta - Acquisition

TUFTS University Ingest Guide

- ISO 20652 Producer-archive interface - Methodology abstract standard (PAIMAS)

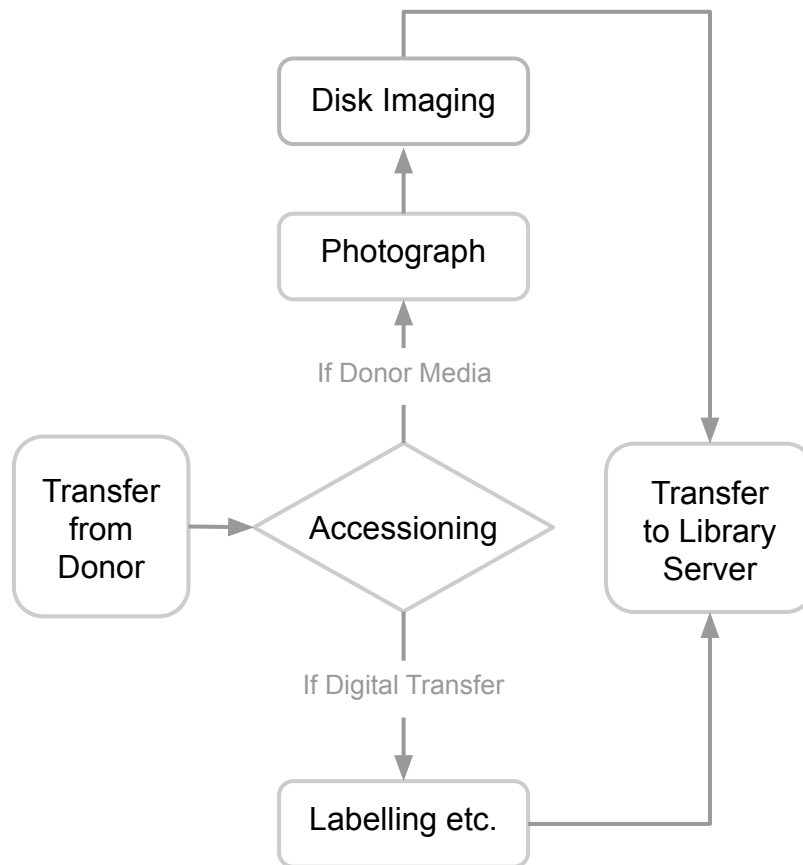
Decision-making and considerations based; not tied to a specific technology



See: <https://dca.lib.tufts.edu/features/nhprc/reports/ingest/index.html>

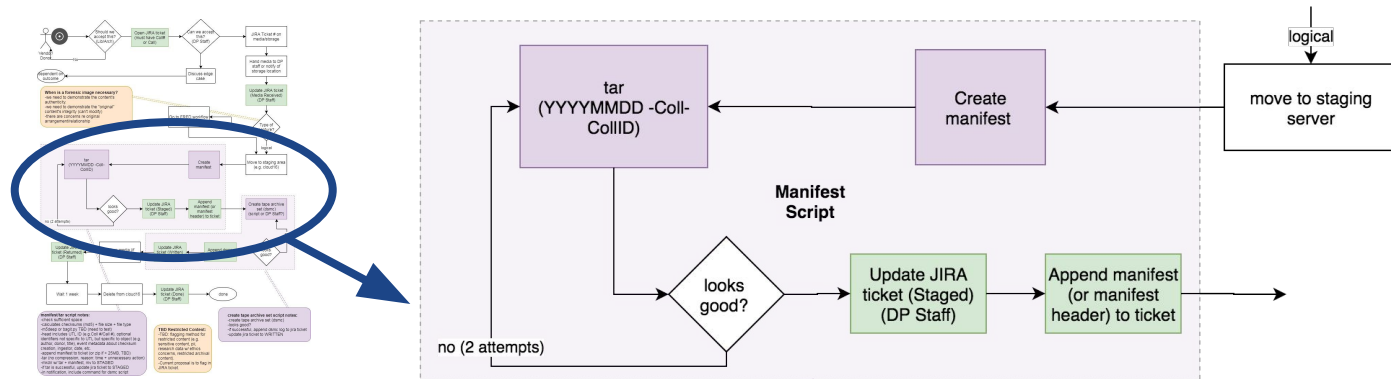
McMaster - Acquisition

- Acquisition considerations same criteria for other archives
- What we are given does determine initial steps after accessioning



Metadata

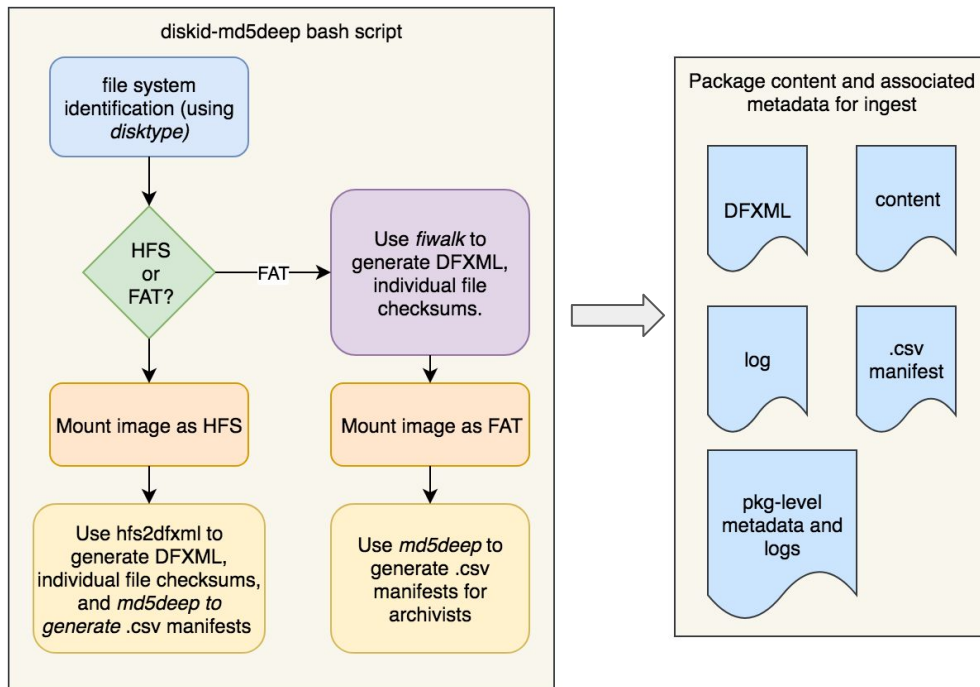
University of Toronto - Metadata (General)



manifest script notes:

- calculates and logs checksums (md5) + file size + file type
- generates event metadata about checksum creation, ingestor, date, etc.
- append manifest to ticket (if <20MB)
- tar (no compression)
- mkdir w/ tar + manifest, YYYYMMDD-LIB-Desc-DISK#
- if successful, update ticket and append manifest

University of Toronto - Metadata (Media)



Note: if disk images are not being kept, manifests are simply created from extracted content rather than disk images.

UK Non-Print Legal Deposit - Metadata

- Quality and type of descriptive metadata depends on what the publishers create and provide as well as standards used
- Descriptive records created in BL's Aleph catalogue (eBooks, eJournals, notated music, and geospatial datasets)
- UK Legal Deposit Web Archive do not have individual Aleph records
- Metadata for all items ingested held in METS file for an item package as well as in Metadata Extension Repository (MER) and the Metadata Database (MDDb)
- Records harvested by the LDLs via OAI-PMH or FTP (for older workflows)



Bodleian Libraries
UNIVERSITY OF OXFORD



LLYFRGELL GENEDLAETHOL CYMRU
THE NATIONAL LIBRARY OF WALES



Coláiste na Tríonóide, Baile Átha Cliath
Trinity College Dublin
Oibcos Átha Cliath | The University of Dublin

University of Alberta - Metadata

Hybrid analog-digital system for descriptive & administrative metadata:

- In-house accession reg; RAD for full finding aid
- Database record of accessions
- Hardcopy accession file with donation information
- Digital and hard copy donations go through same accession registration process
- Digital files, like physical file folders, labelled with accession & file number

University of Alberta - Metadata

Donor required to:

- Confirm & finalized file listing (PAIMAS requirement)
- List known rights issues (copyright, privacy, restrictions, etc.)

SIP/AIP/DIP division

- By accession
- Separate Open vs. Restricted package for accession

Permafrost - Metadata

- Subscribers encouraged to map descriptive metadata to simple DublinCore for ingest to Archivematica
 - Transfer-level metadata can be added through interface; item-level metadata is imported via CSV
 - MakeCSV tool can be used to create CSV template with file name listings for easier import
- Use of PREMIS rights metadata entry/import also encouraged

Subscriber's local system



Optionally installed locally:

- Whatever hardware/software required for local data capture/creation/storage
- MakeCSV Tool
- BagIt-compliant tool (Exactly, Bagger, etc.)

User packages
up data and
metadata for
preservation

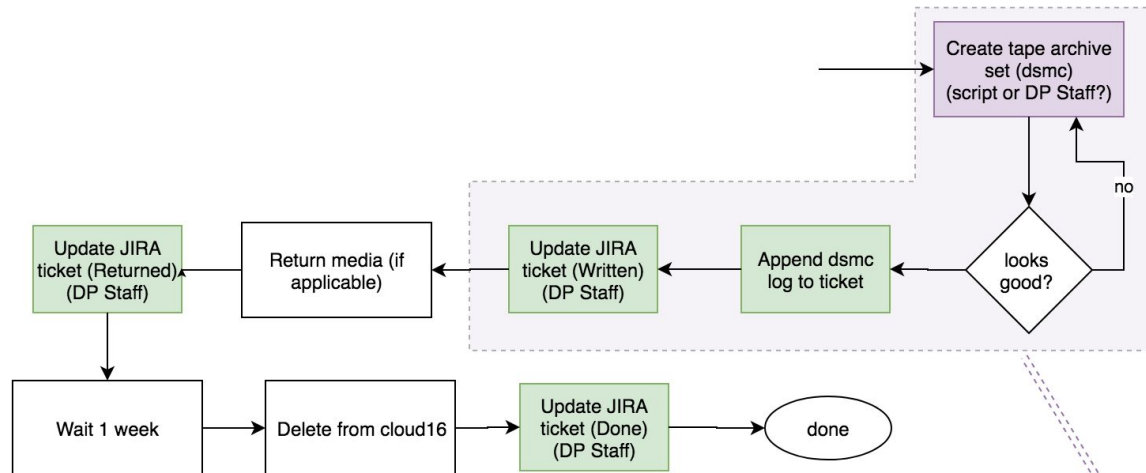
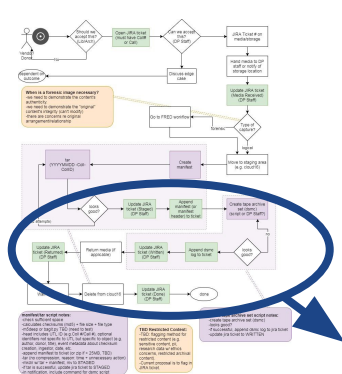
McMaster - Metadata

Combination of RAD finding aid and repository metadata needs

- All digital files linked to accession number to pair with physical records
- Descriptive metadata from file names, media labels, accession records etc.
- PREMIS rights metadata
- Written in .CSV mapped to DublinCore

SIP Ingest

University of Toronto - Ingest (General)



TBD Restricted Content:

TBD: flagging method for restricted content (e.g. sensitive content, pii, research data w/ ethics concerns, restricted archival content).

create tape archive set script notes:

- create tape archive set (run tsm command)
- if successful, append log to jira ticket
- update jira ticket to WRITTEN

UK NPLD Legal Deposit - SIP Ingest

Pre-Ingest

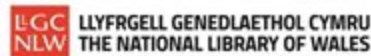
- Steps depend on content type and source
- Content virus checked
- Content packaged if not received this way from publisher (e.g. for eBooks, a metadata file, content file, and often .jpg of cover)
- Metadata transformation
- Creation of METS file
- Persistent identifier assigned to each object (Archival Resource Key / ARK)

Strategic Ingest (common ingest workflow)

- METS data added to Metadata Extension Repository (MER)
- Record created in Aleph
- Metadata from Aleph and MER combined to create an aggregated record for each item
 - Added to Data Warehouse for management information purposes
 - Published to OAI-PMH so records can be harvested by other Legal Deposit Libraries
 - Published to Primo to enable discovery at the BL
- DOMID (internal identifier) assigned to each object



Bodleian Libraries
UNIVERSITY OF OXFORD



University of Alberta - SIP

Records first go into a staging area for Archivist intervention (relabelling, metadata, etc.)

Divided by accession & subdivided by open vs restricted

Original files & metadata in SIPs, as appropriate

Permafrost - SIP

- User packages transfers on local systems containing
 - Original materials to be preserved
 - Descriptive metadata (if imported)
 - Rights metadata (if imported)
 - Contextual documentation (if available)

Subscriber's local system



Optionally installed locally:

- Whatever hardware/software required for local data capture/creation/storage
- MakeCSV Tool
- BagIt-compliant tool (Exactly, Bagger, etc.)



OwnCloud folder hosted on
OLRC

User packages
up data and
metadata for
preservation

User drags and drops to
transfer source on OLRC; this
is synced to a local folder on
the Archivematica Virtual
Machine

- Transfers are (optionally) bagged and dragged and dropped to OLRC transfer point staging area
 - This staging area will (likely) use OwnCloud - a Dropbox-like application hosted on the OLRC (currently in testing!)
 - The OwnCloud folder is synced to a local folder on the Archivematica VM to reduce bottleneck at transfer point

McMaster - SIP

SIP to contain:

- Original files
- Rights information
- Photographs of donor media (when present)
- Metadata

Grouped by accession, rights, and archival hierarchy (sometimes)

Preservation

UK Non-Print Legal Deposit - Preservation

As part of either pre-ingest or ingest workflow:

- Checksum creation and/or validation
- File format identification, characterisation, and validation (JHOVE for PDFs and EpubCheck for EPUBs; METS validated against BL METS profile)
- Digital signatures created for every object
 - Later validated with Object Authenticity Checker (OAC) tool
- All objects (individual digital files) ingested and created during ingest (e.g. METS file) preserved within store and replicated four times

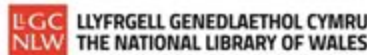
UK Non-Print Legal Deposit - Preservation

What else helps ensure immediate and long-term preservation?

- Undertaking an audit/assessment of technical infrastructure, staffing, and governance for preservation
- Preservation policy and implementation plan
- Preservation planning (How are changes that could impact the preservation of and access to content monitored? What information informs preservation actions?)
- Collection profiles (e.g. high-level information on why content is being preserved and main preservation requirements)
- Providing access to collections and listening to user feedback



Bodleian Libraries
UNIVERSITY OF OXFORD



Coláiste na Tríonóide, Baile Átha Cliath
Trinity College Dublin
Ollscoil Átha Cliath | The University of Dublin

University of Alberta - Preservation

SIPs are bagged along with technical & structural metadata

- Currently, custom made “lightweight AIP” process, looking to move to “full AIP”

2-tiered preservation:

1. Bit level

- Guaranteed for all deposits
- “Preservation storage”

University of Alberta - Preservation

2. Full preservation

- For records that are legally required to be maintained
- Renderable on an ongoing basis
- Combination of normalization (file format registry based on LOC registry), migration, and emulation as necessary based on content & format type
- Will try our best for as much as possible. Higher likelihood for common formats

University of Alberta - Preservation

AIP contains:

- Original files
- Normalized and/or migrated records, as applicable
- Metadata
 - Descriptive
 - Administrative
 - Technical
 - Structural

Permafrost - Preservation

- Preservers process package through “transfer” stage workflow in Archivematica to create SIP
 - Various functions performed: file identification, characterization, validation
 - Processing configuration can be adapted as needed; automated if desired
- Preservers optionally normalize files
 - Work with individuals to assess normalization options
 - Not all ‘out of the box’ Archivematica normalization rules are good for everyone



Dedicated Virtual Machine
running Archivematica

User processes materials
through Archivematica
workflow

McMaster - Preservation

AIP created in Archivematica always has:

- Original files
- Metadata

Likely contains:

- Normalized files (if needed/available)
- Photographs of original media (planned, not yet active)

AIP Storage

UK NPLD Legal Deposit - AIP storage

- NPLD objects ingested at BL Boston Spa
- Ingested objects replicated to four geographically separate nodes: British Library Boston Spa and St Pancras, National Library of Scotland, National Library of Wales
- Integrity checking carried out on each of the nodes using OAC
- Self-healing from another node when an object is not found or is corrupt
 - Corrupt objects quarantined



University of Alberta - Storage

“Preservation storage”:

- Mirrored, triplicate backup storage
- “Self healing”: ongoing integrity checking via checksums
- Physically separated secure server rooms
- Multi-layer fire walls

Currently using OpenStack Swift

Scholars Portal - Storage

- Subscriber stores AIP on designated, institution-specific account on OLRC
- Copies are replicated to at least 3 of 5 sites in Ontario: Toronto, Queens, Ottawa, York & Guelph
- Based on open source OpenStack Swift platform
- Self-healing through checksums
- Uses ORION high speed network
- Data centres housed in secure centres at universities; uses SSL encryption
- Next up: better management layer for AIPs in OLRC through project Canopus



McMaster - Storage

I do what they do.

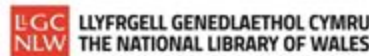
DIP & Access

UK NPLD Legal Deposit - DIP and Access

- Preservation and access files the same for eBooks (EPUB), eJournals (PDF), and notated music (PDF)
 - Calibre viewer used for EPUBs and Sumatra PDF for PDFs
- Access to preserved websites provided via UK Web Archive
- Access to geospatial datasets provided via theMapCloud platform
- Ex Libris Primo used for Discovery at BL
- Records harvested by other LDLs via OAI-PMH
 - Records added to catalogues at each LDL to enable search, discovery, and access



Bodleian Libraries
UNIVERSITY OF OXFORD



Coláiste na Tríonóide, Baile Átha Cliath
Trinity College Dublin
Ollscoil Átha Cliath | The University of Dublin

UK NPLD Legal Deposit - DIP and Access

- Single concurrent access per item per Library
- Users can only access content on library terminals
- No digital copies can be removed from the reading room
- Printing allowed in accordance with Copyright Regulations
- No digital sharing or screen shots
- No interlibrary loan unless it is between one legal deposit library and another
- Cannot text data mine



Non-Print (Electronic) Legal Deposit Access Policy

The British Library provides access to this item on certain conditions

We expect everyone to **read and accept** the following guidelines on appropriate use of non-print legal deposit material before they begin to view material using this system. All usage will be monitored. **Infringement** of these guidelines may lead to suspension of your reader pass and **legal proceedings** by the rights owner.

Permitted use according to the Regulations

This material has been supplied to the British Library under the Non-Print Legal Deposit Regulations and as such its use is limited under these regulations.

- a) Digital copies of these items are not permitted but printing is permitted in accordance with Copyright Regulations [as described online](#) and on posters around the Reading Room.
- b) Text and data mining is not permitted on this content.

For non-commercial research or private study of born-digital sheet music, fair dealing rules apply. The Music Publisher's Association guidelines advises that anything copied should be less than a performable piece to be used for study, and should not be intended or used for performance. A licence from the MPA will be needed for copying for performance.

Non-Print Legal Deposit items can only be accessed by a single user at any one time therefore as soon as you have finished with an item please close the browser to allow another reader to access the content.

- c) Full details can be found in the Legal Deposit web pages:
www.bl.uk/aboutus/legaldeposit/introduction/index.html

If you intend to **print this item** the **Accept** button counts as an electronic signature agreeing to the following

I declare that:

- a) I will not use the copy except for the purposes of research for non-commercial purpose; private study; criticism or review or reporting current events; parliamentary or judicial proceedings; or a Royal Commission or statutory inquiry.
- b) In relation to a copy of relevant material required for the purposes of non-commercial research or private study:
 - i I have not previously been supplied with a copy of the same material by you or another deposit library; and
 - ii To the best of my knowledge, no person with whom I work or study has made or intends to make, at or about the same time as this request, a request for substantially the same material for substantially the same purpose.

I understand that, if the declaration above is false in a material particular, the printed copy supplied will be an infringing copy and that I shall be liable for infringement of the Copyright.

ACCEPT

DECLINE

Please close the browser tab when you have finished viewing this item



Search for specific URL (e.g. [www.bl.uk](#)) or any word or phrase. [Search tips](#)

The UK Web Archive (UKWA) collects millions of websites each year and preserves them for future generations.

UKWA collects on behalf of the UK Legal Deposit Libraries - The British Library, National Library of Scotland, National Library of Wales, Bodleian Libraries, Cambridge University Libraries and Trinity College, Dublin.

[About us](#)

Highlights from the Special Collections

Special Collections are groups of websites brought together on a particular theme by librarians, curators and other specialists, often working in collaboration with key organisations in the field.

[View all special collections](#)



Layers



Metadata



Print

- ☐ Postcode Sector (S16) NE England
- ☐ Postcode Sector (S17) Scottish Central Belt
- ☐ Postcode Sector (S18) Scottish Central Belt West
- ☐ Postcode Sector (S19) Highlands & Islands
- ☐ Postcode Sector (S20) Angus & Aberdeenshire
- ☐ UK NHS Strategic Health Authorities

GeoInformation Group**UK Buildings****Legend**

UK Buildings Polygons 2016

- Residential
- Non Residential Building
- Mixed Residential and Non Residential B
- Unknown

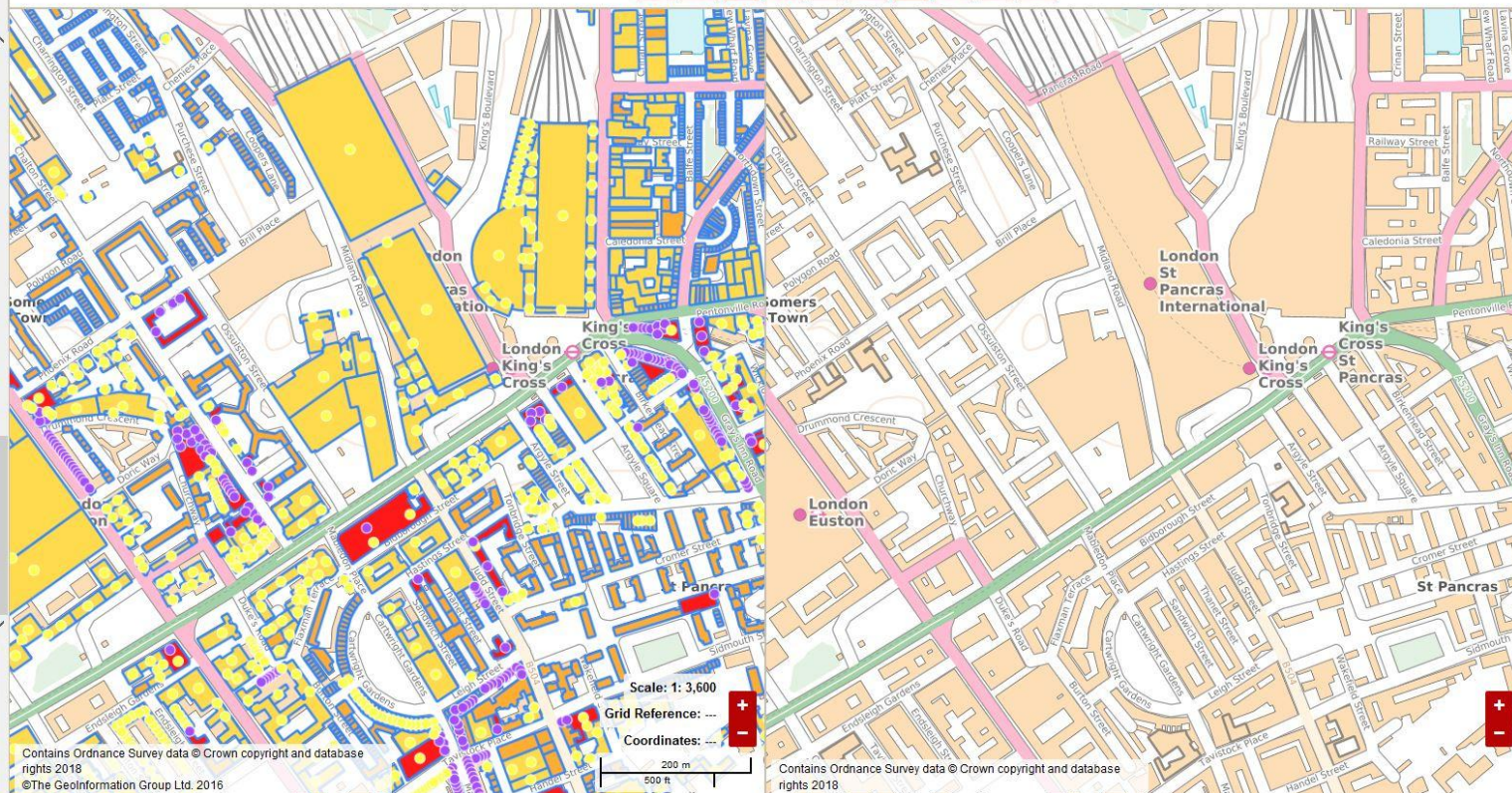
UK Buildings Points 2016

- Non Residential
- Mixed Residential and Non Residential

- ☐ UK Land
- ☐ UK Map (London only)

+ Ordnance Survey

UK Buildings 2016



Contains Ordnance Survey data © Crown copyright and database rights 2018
©The GeoInformation Group Ltd. 2016

Contains Ordnance Survey data © Crown copyright and database rights 2018

University of Alberta - DIP & Access

Open access DIP (free of rights concerns, including privacy, restrictions, copyright):

- Metadata available on Archives database (AtoM)
- DIP stored on a library server & made available through Archives database
- Routine Disclosure for university records

Restricted DIP (subject to rights concerns):

- Metadata available on Archives database (AtoM), unless otherwise restricted
- Users request access from Archivist. If granted, DIP created on demand
- Secure, timed link provided by Archivist to user

Scholars Portal - DIP & Access

- Optional to create DIP and store on OLRC
- Not currently linking to/hosting access platforms because use cases vary so much between members
 - But most are using locally-hosted AtoM instances for archival materials
 - Testing automatic AtoM DIP upload from hosted Archivematica to local AtoM
- Other in-progress workflows actually start with a DIP and work backwards to create an AIP
 - Islandora
 - Dataverse



DIP to OLRC, AtoM, etc.

Fixity,
storage
stats
reported
monthly

AIP & DIP stored
stored on OLRC or DIP
transferred elsewhere

McMaster - DIP & Access

Open DIP:

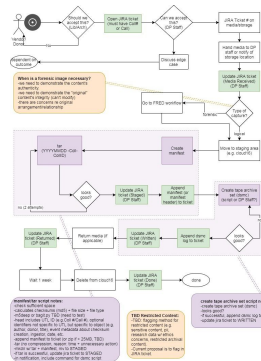
- Metadata available in online finding aid or AtoM
- DIP stored on a library server & made available through Islandora, AtoM, or (possibly) ePADD

Restricted DIP:

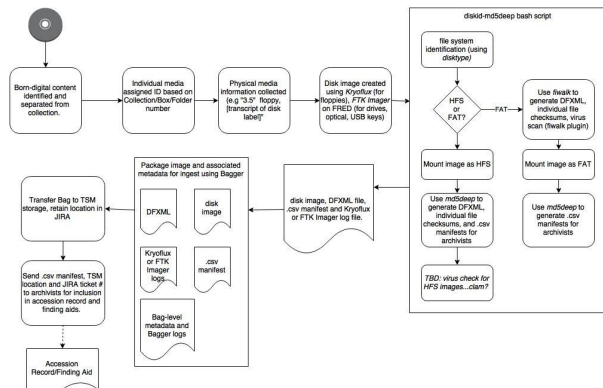
- Metadata available in online finding aid or AtoM (as appropriate)
- DIP stored on library server, accessible through separate Islandora instance on a terminal in the reading room with no internet access

Workflow Diagrams

University of Toronto - Intake Workflow Links



<http://uoft.me/gen-workflow>



<http://uoft.me/media-workflow>

University of Alberta - Workflow

WHY

Stewardship

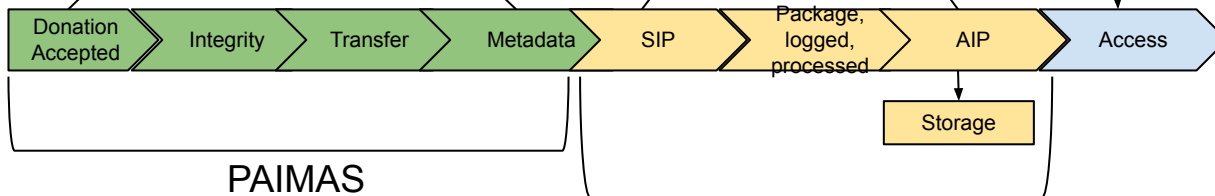
WHAT

Acquisition

Preservation

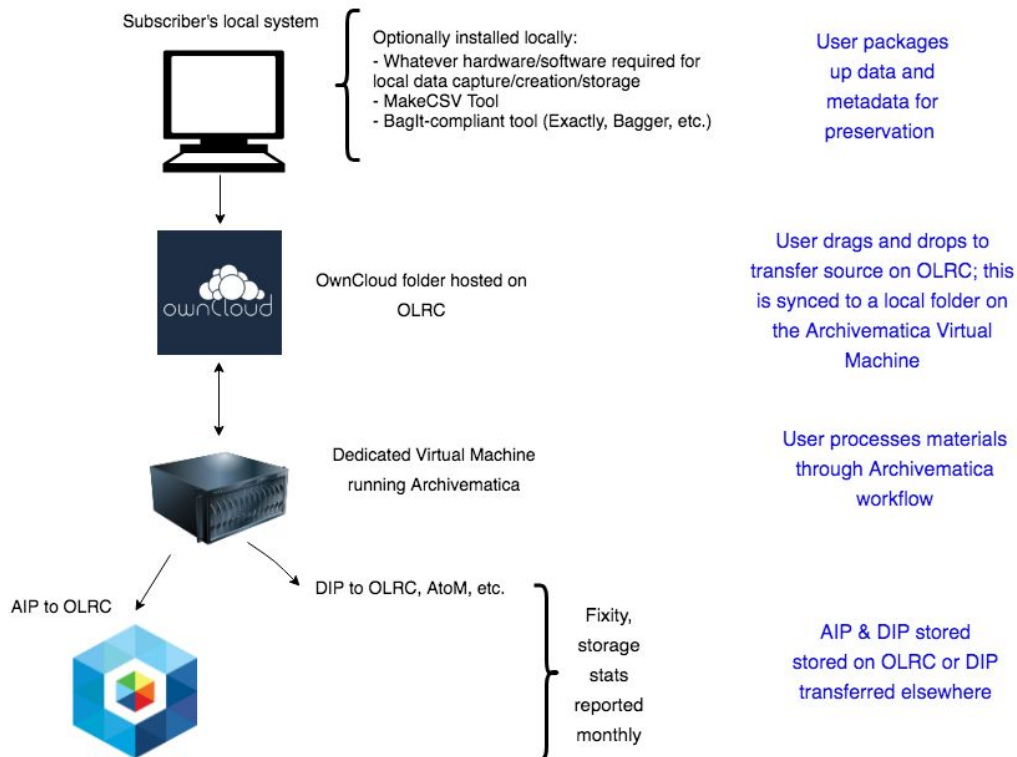
Research and
Access*

HOW

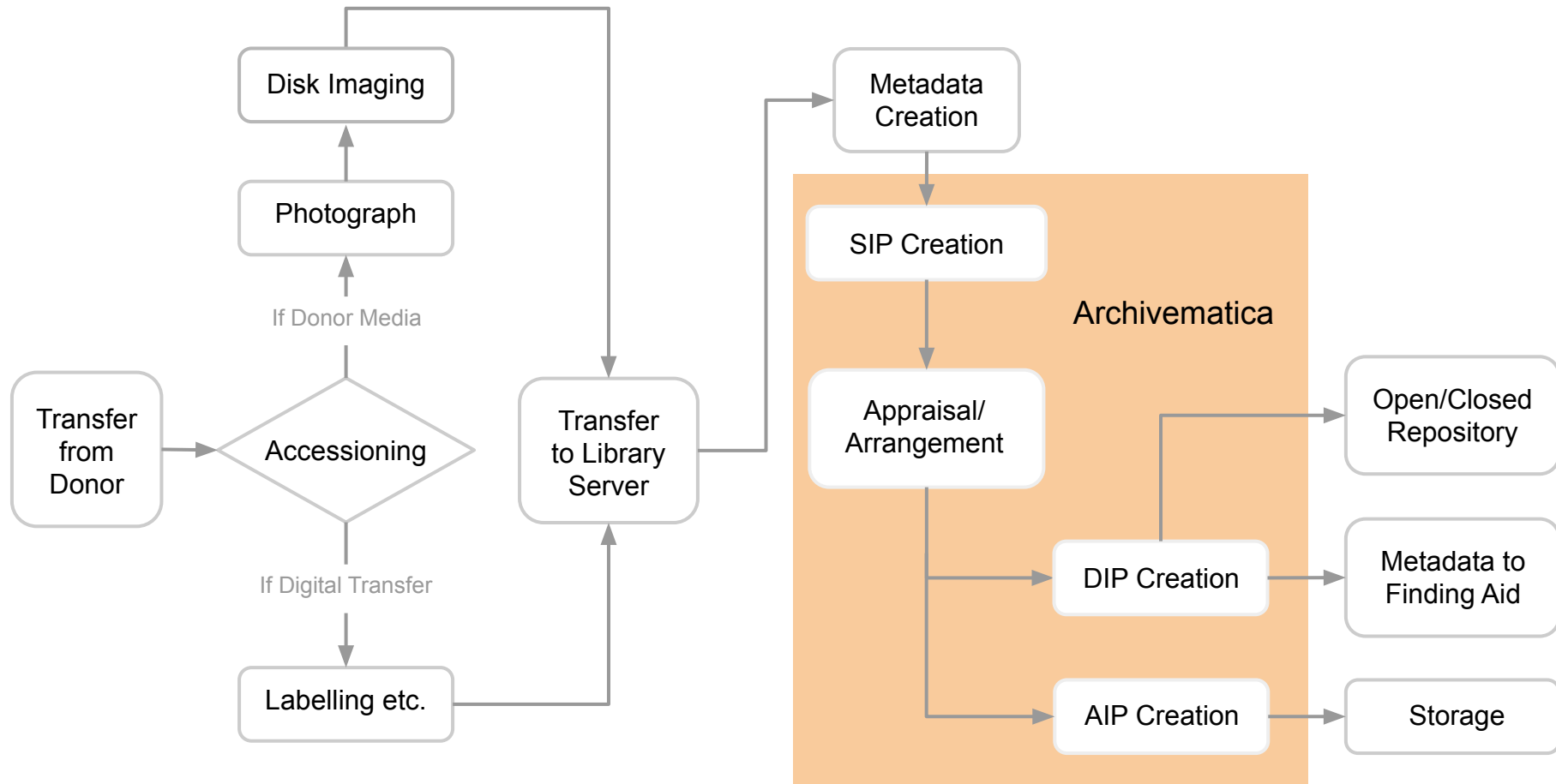


OAIS

Permafrost - Workflow



McMaster - Workflow



Thank you for joining us!

Questions?

Jess Whyte - jessica.whyte@utoronto.ca

Caylin Smith - caylin.smith@bl.uk

Krista Jamieson - kjamieso@ualberta.ca

Grant Hurley - grant.hurley@utoronto.ca

Bridget Whittle - whittle@mcmaster.ca