

The Data Archivist

The archivist's role in data management
and preservation

Sara Allain & Sarah Romkey | Artefactual Systems, Inc.

May 26, 2016 | TCDL 2016, Austin, Texas



Today's talking points

The role of the archivist in research data management

Basic intro to Archivematica

Three case studies:

1. Ontario Council of University Libraries + Dataverse
2. University of York & University of Hull + Hydra
3. Compute Canada + Globus



Archivists

+

RDM



RDM Isn't New

We've been thinking about the role of the library in research data management for several years.



The Digital Preservation Gap

Digital management platforms must adequately preserve data. Domain-specific tools and proprietary formats make this difficult.



Assertions

Research data management is a digital preservation problem.
Archivists are pretty good at digital preservation.



Why Archivematica?

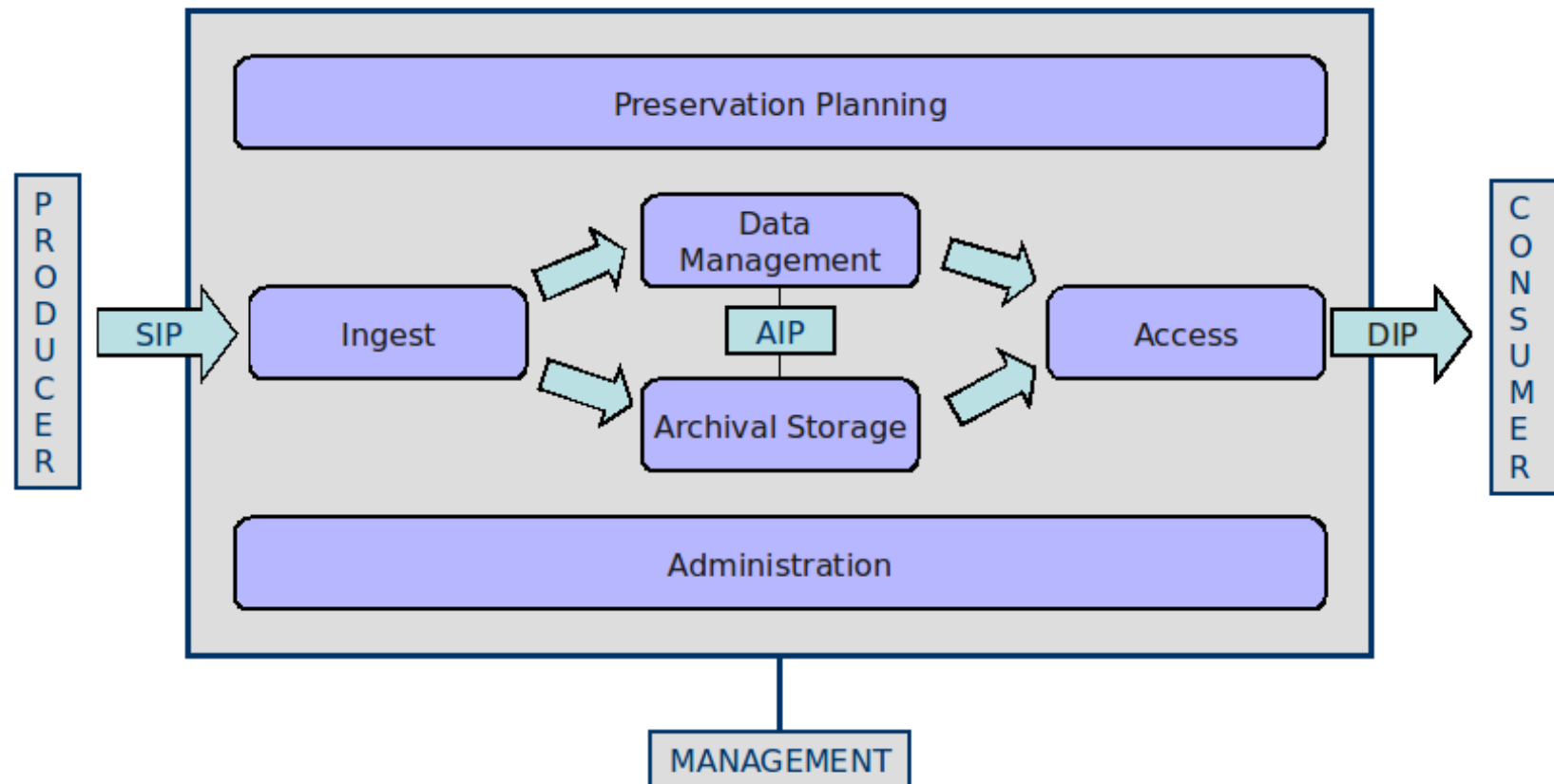


Definition

Web- and standards-based open-source application which allows your institution to preserve long-term access to trustworthy, authentic and reliable digital content.



Open Archival Information System (OAIS) reference model (ISO-STD 14721)






Standard
Type

Transfer name

Accession no.

/home
Browse

Start transfer

Transfer	UUID	Transfer start time	
<div> <div>  test </div> <div> <div> Micro-service: Create SIP from Transfer Micro-service: Complete transfer Micro-service: Examine contents Micro-service: Validation Micro-service: Characterize and extract metadata Micro-service: Update METS.xml document Micro-service: Extract packages Micro-service: Identify file format Micro-service: Clean up names Micro-service: Generate transfer structure report Micro-service: Scan for viruses Micro-service: Quarantine Micro-service: Generate METS.xml document Micro-service: Verify transfer checksums Micro-service: Reformat metadata files Micro-service: Assign file UUIDs and checksums Micro-service: Include default Transfer processingMCP.xml Micro-service: Verify transfer compliance Micro-service: Rename with transfer UUID Micro-service: Approve transfer </div> </div> </div>	6c371e09-57c5-4225-b1c2-b2cfc7cff883	2016-05-25 09:23	<div>   </div>



Any

Keyword

Search transfer backlog

[Add New](#)

originals

Hide






View File

arrange

Delete
Create SIP
Edit Metadata
Add Directory

Submission Information Package	UUID	Ingest start time	
<div> <div>test</div> <div> <div>Micro-service: Upload DIP</div> <div>Micro-service: Store AIP</div> <div>Micro-service: Prepare AIP</div> <div>Micro-service: Prepare DIP</div> <div>Micro-service: Generate AIP METS</div> <div>Micro-service: Process metadata directory</div> <div>Micro-service: Verify checksums</div> <div>Micro-service: Process submission documentation</div> <div>Micro-service: Transcribe SIP contents</div> <div>Micro-service: Add final metadata</div> <div>Micro-service: Normalize</div> <div>Micro-service: Process manually normalized files</div> <div>Micro-service: Clean up names</div> <div>Micro-service: Verify transfer compliance</div> <div>Micro-service: Remove cache files</div> <div>Micro-service: Include default SIP processingMCP.xml</div> <div>Micro-service: Verify SIP compliance</div> <div>Micro-service: Rename SIP directory with SIP UUID</div> </div> </div>	4bfdc9be-5abe-45bb-8674-8c4340c6c740	2016-05-25 09:28	<div> <div></div> <div></div> </div>



▼ Micro-service: Normalize		
Set file permissions	Completed successfully	
Move to processing directory	Completed successfully	
Approve normalization [?]	Completed successfully	 
Set file permissions	Completed successfully	
Move to approve normalization directory	Completed successfully	
Remove files without linking information (failed normalization artifacts etc.)	Completed successfully	
Normalize for preservation	Completed successfully	
Normalize for access	Completed successfully	
Normalize for thumbnails	Completed successfully	
Create thumbnails directory	Completed successfully	
Create DIP directory	Completed successfully	
Move to processing directory	Completed successfully	
Normalize [?]	Completed successfully	
Resume after normalization file identification tool selected.	Completed successfully	
Identify file format	Completed successfully	
Select pre-normalize file format identification command	Completed successfully	
Move to select file ID tool	Completed successfully	
Grant normalization options for no pre-existing DIP	Completed successfully	
Set remove preservation and access normalized files to renormalize link.	Completed successfully	
Check for Access directory	Completed successfully	
Check for Service directory	Completed successfully	
Identify manually normalized files	Completed successfully	



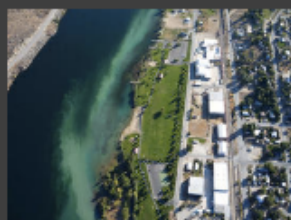
Holdings

Quick search

▼ Collection TCDL1 - TCDL test

Item [Landing zone.jpg](#)Item [MARBLES.TGA](#)

Collection TCDL1 - TCDL test



Landing zone.jpg

Identity area



Reference code	TCDL1
Title	TCDL test
Date(s)	• 2016-5-25 - 2016-5-26 (Creation)
Level of description	Collection
Extent and medium	Extent and medium

Context area



Name of creator	Example creator
-----------------	---------------------------------

Access points



Name access points	• Example creator (Creator)
--------------------	---

[Reports](#)

Import

[XML](#) [CSV](#)

Export

[Dublin Core 1.1](#)[XML](#) [EAD 2002 XML](#)

Finding aid

[Generate](#) [Status: Unknown](#)Related people and
organizations[Example creator](#)
(Creator)

All Packages

Packages are Transfers, SIPs, DIPs and AIPs uploaded to a Location managed by the storage service.

[View recovery requests](#) | [View delete requests](#)

Show 10 entries

Search: 4bf

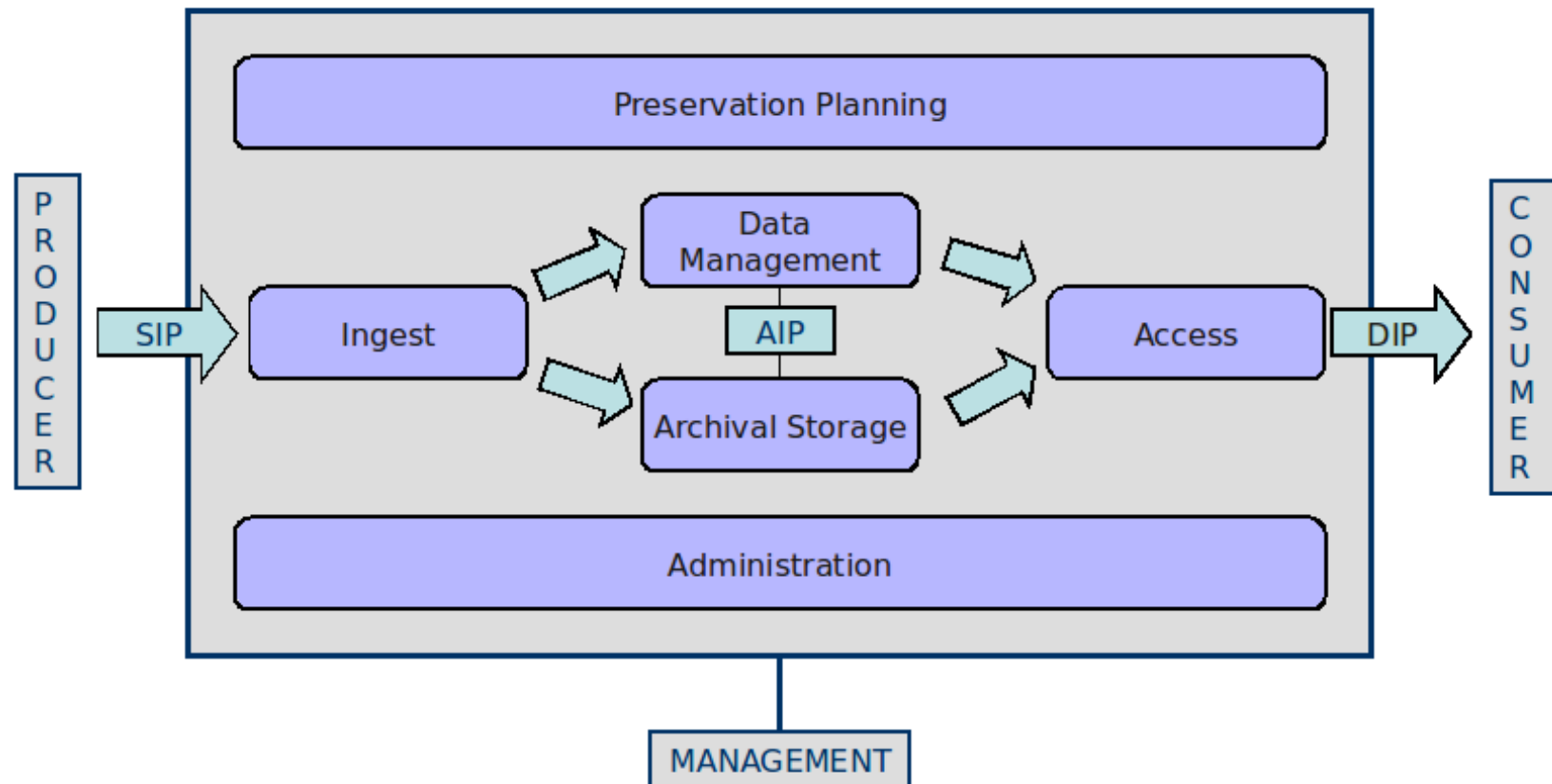
UUID	Description	Originating Pipeline	Current Location	Size	Type	Pointer File	Status	Actions
4bfdc9be-5abe-45bb-8674-8c4340c6c740	None	Archivematica on apricotjelly.archivematica.org (b66b66ea-eb05-4fc8-a24b-d8bd4d584375)	/var/archivematica/sharedDirectory/www/AIPsStore/4bfd/c9be/5abe/45bb/8674/8c43/40c6/c740/test-4bfdc9be-5abe-45bb-8674-8c4340c6c740.7z	20.3 MB	AIP	Pointer File	Uploaded (Update Status)	Download Re-ingest
d164244c-4d33-41cb-b3ad-ee6aa9e585c7	None	Archivematica on apricotjelly.archivematica.org (b66b66ea-eb05-4fc8-a24b-d8bd4d584375)	/var/archivematica/sharedDirectory/www/DIPsStore/d164/244c/4d33/41cb/b3ad/ee6a/a9e5/85c7/test-4bfdc9be-5abe-45bb-8674-8c4340c6c740	1.5 MB	DIP	None	Uploaded (Update Status)	Download

Showing 1 to 2 of 2 entries (filtered from 45 total entries)

[Previous](#) [Next](#)



Open Archival Information System (OAIS) reference model (ISO-STD 14721)



So... Why Archivematica?

Based on standards and best practices

Format and repository agnostic

Small enough to run on a laptop

Robust enough to handle petabytes of data

Modular

Free and open source

Familiar



Archivematica is for Archivists

It was built around archival standards, using archival terminology, and it's meant to anticipate archival digital preservation workflows. (Of course, everyone's welcome to use it!)

Luckily, since RDM is a digital preservation problem, it's well suited to RDM workflows as well.



York/Hu11

+

Hydra

Case Study 1



Research Data Spring

Jisc-funded projects aimed at encouraging tool and workflow development to tackle various aspects of research data management.

Available project funding was anywhere from £250k to £1m.



Research Data Spring

Jisc-funded projects aimed at encouraging tool and workflow development to tackle various aspects of research data management.

Available project funding was anywhere from £250k to £1m.



Research Data Spring

York and Hull were successful at obtaining funding for all three phases of the project.

Goal was to take advantage of Archivematica's modularity to integrate Archivematica into a research data management architecture that would include other applications for deposit, management, etc.



York & Hull at the Outset

Established Hydra-based institutional repository, but no digital preservation capacity.

Wanted to be able to offer assured long-term preservation to faculty members.



Archivematica Falls Short!

After Phase 1 (testing), the archivists at York and Hull identified several areas where Archivematica was not sufficient to meet their RDM needs.

They applied for Phase 2 funding to begin developing solutions for the identified problems.



Winter of ~~Our Discontent~~ Development

Five deliverables:

- On demand automated DIP generation
- METS parsing
- Generic search REST API
- Multiple checksum algorithms
- Handle unidentified files

Disclaimer: York and Hull are lovely to work with! But who can resist a Shakespeare joke?



Deploy! Deploy!

York and Hull successfully applied for Phase 3 funding to build a proof-of-concept platform, making use of the deliverables to integrate Archivematica with Hydra.

Meanwhile, Artefactual is currently bundling the new features into the 1.5 and 1.6 releases of Archivematica.



OCUL + Dataverse

Case Study 2



Dataverse at OCUL

Open source repository platform developed at Harvard.

Ontario Council of University Libraries' tech branch, Scholars Portal, hosts a Dataverse instance that is available to academics at Ontario's 21 universities.



Deposit and Access Reign

Dataverse excels as a deposit and access system, but has limited digital preservation functionality.

Goal of the project was to let users deposit content through Dataverse, running Archivematica preservation tasks in the background.

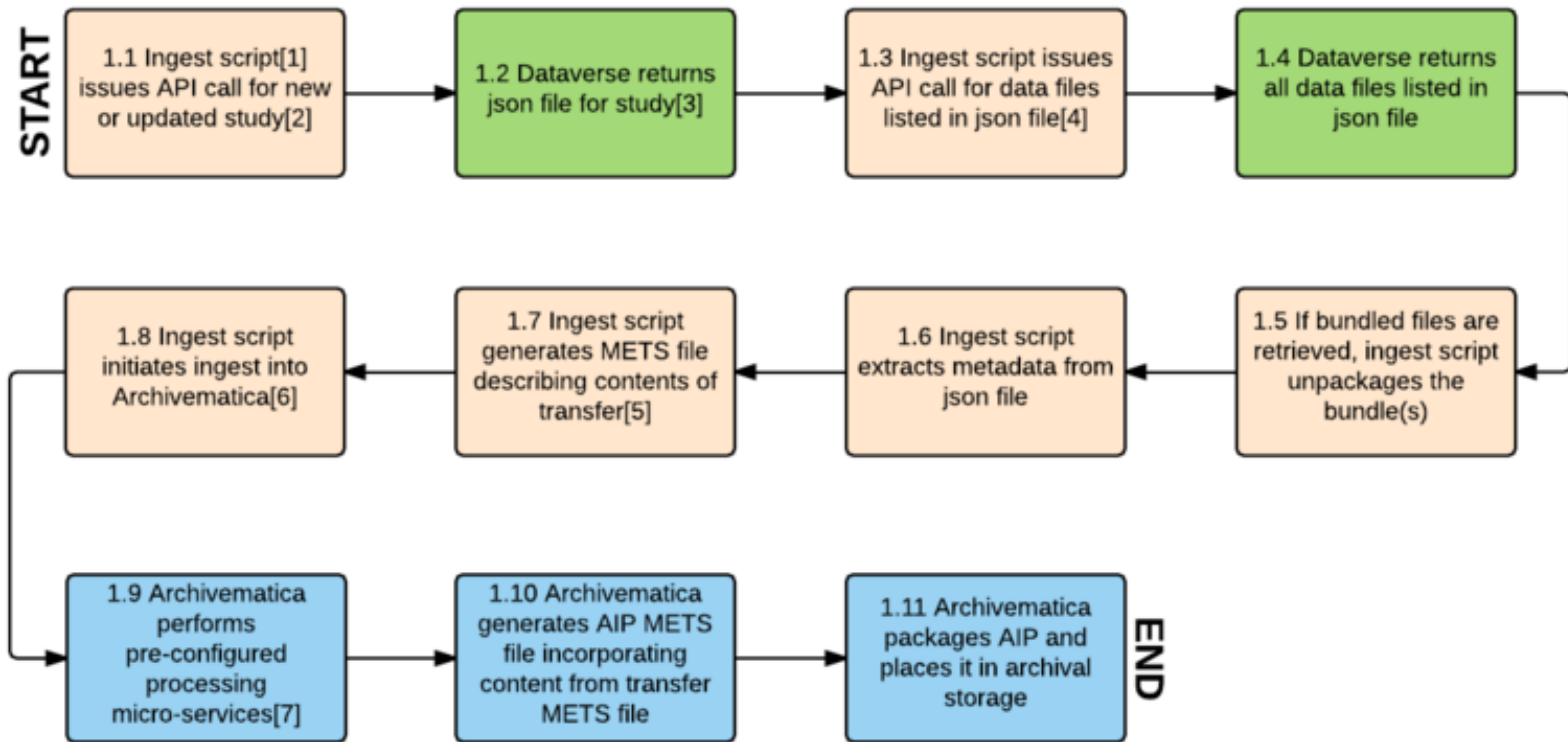
Important: users can deposit content over time, rather than all at once!



Automate It!

The integration makes use of Automation Tools, an Archivematica library that facilitates requests for updated information from Dataverse's API. An ingest script was also developed to manage ingest tasks.





Orange: Automation Tools

Green: Dataverse

Blue: Archivematica



An Experiment

The Dataverse integration project resulted in a proof of concept workflow that isn't currently scheduled for release. However, it's available as a separate public branch of the project on Github.

At some point in the future, we would love to generalize the code and make it available in a public release.



Compute Canada + Globus

Case Study 3



Compute Canada

A national, non-profit organization that provides high performance research computing resources for 70 institutions and 10,000+ researchers.

Compute Canada uses Globus' Transfer Service and Publication Service tools to store and provide access to research data.

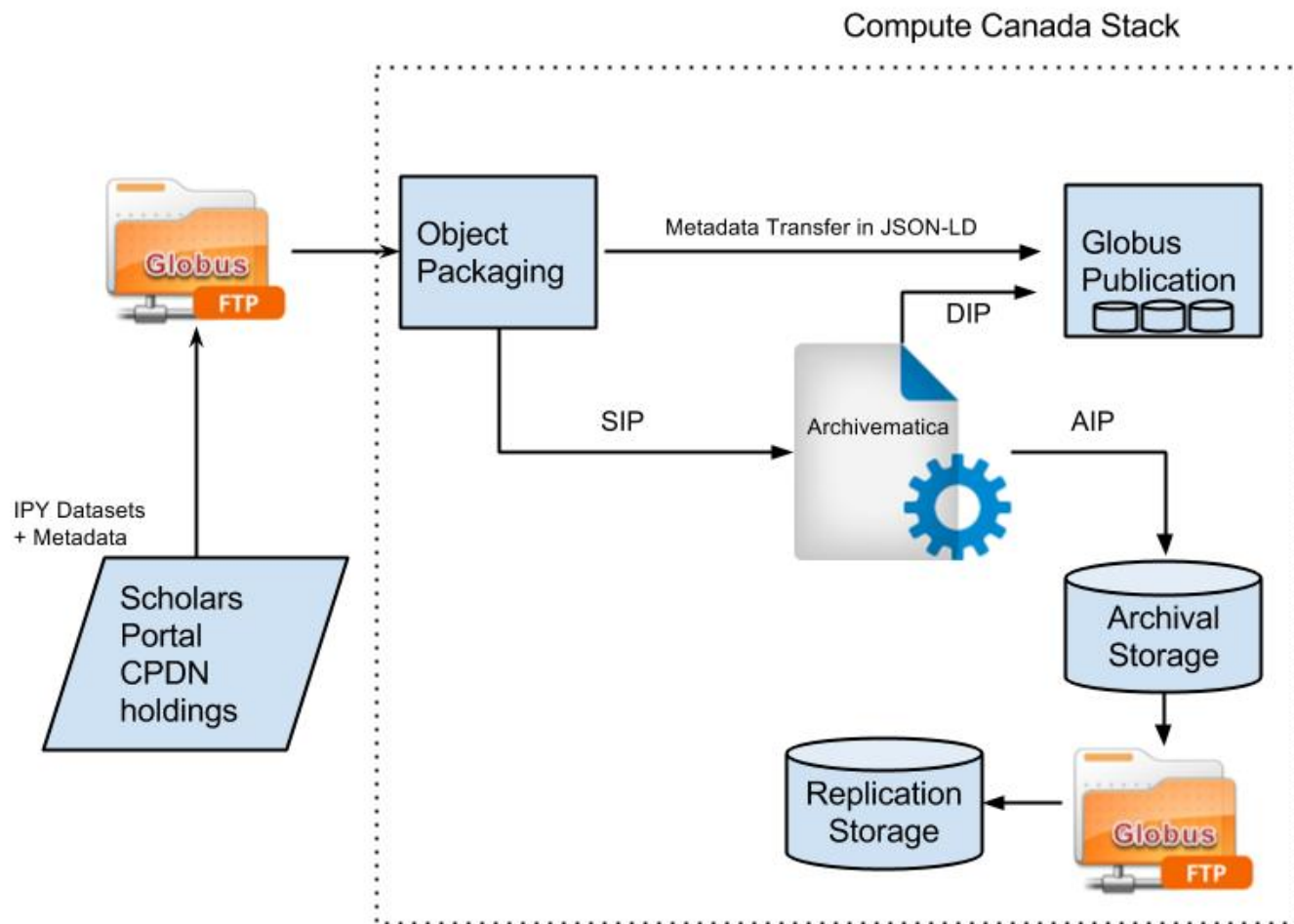


Canadian Polar Data Network Pilot

Scholars Portal holds terabytes of climate data from the CPDN. This corpus was used to pilot an integration where Archivemata acts as a bridge between the Globus Transfer and Publication Services and Compute Canada datastores.



CPDN RDC Federated Pilot Diagram Version 4



Another Experiment

This proof of concept is also not scheduled for release. We're working on getting it into a separate public branch of the Archivemata project on Github.



Archivists

+

RDM



Get In Touch!

Twitter: [@archivalistic](#) | [@archivematica](#)

Email: sallain@artefactual.com or info@artefactual.com

This presentation: bit.do/data-archivist

