

Pre-Ingest processing for digital archival records

Heather Tompkins
Senior Project Officer
Digital Integration
Library and Archives Canada (LAC)

DISCOVER. UNDERSTAND. CONNECT.



Library and Archives
Canada

Bibliothèque et Archives
Canada

Canada

Library and Archives Canada

LAC is both a national library and a national archives

Broad 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, contributing to the cultural, social and economic advancement of Canada as a free and democratic society
- to facilitate in Canada co-operation among communities involved in the acquisition, preservation and diffusion of knowledge
- to serve as the continuing memory of the Government of Canada and its institutions

What is Pre-Ingest?

Pre-Ingest = Technical Appraisal

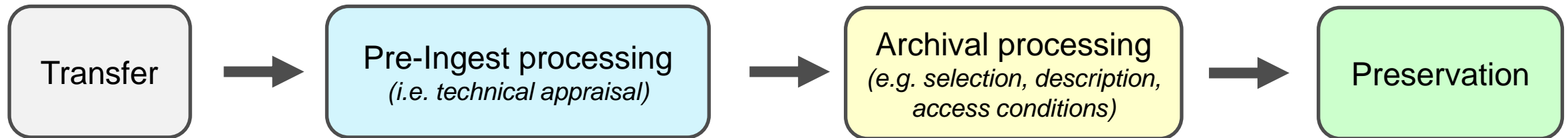
- Performed on digital archival records
- First stage of processing transferred digital records at LAC
 - Occurs post-transfer but prior to content being under the care of LAC's Digital Preservation section
 - Occurs before selection, arrangement & description by archival staff

Why do we need Pre-Ingest?

- Two key reasons:
 - Not every transfer is ready for preservation as-is
 - We may not want to keep everything transferred to us
- Context of government vs private transfers
- Pre-Ingest allows us to:
 - Know the nature of transferred digital records
 - Understand where digital preservation requirements are not met
 - Make informed decisions about what to preserve and how to preserve it

Where does Pre-Ingest fit?

A high-level overview of where Pre-Ingest processing falls within the overall processing workflow:



How do we do Pre-Ingest?

Pre-Ingest Workflow

What is pre-ingest?

Pre-ingest is the technological review of digital records transferred to LAC. The goal of pre-ingest is to ensure that the digital records are in a format that can be preserved using the software tools to automate the process. The goal of pre-ingest is to ensure that the digital records are in a format that can be preserved using the software tools to automate the process. The goal of pre-ingest is to ensure that the digital records are in a format that can be preserved using the software tools to automate the process.

What are the major tasks are for pre-ingest?

- Weed any digital records that should not have transferred (e.g. configuration, developmental, temporary, software files etc)
- Identify file format or other² issues that need addressing by archival staff.

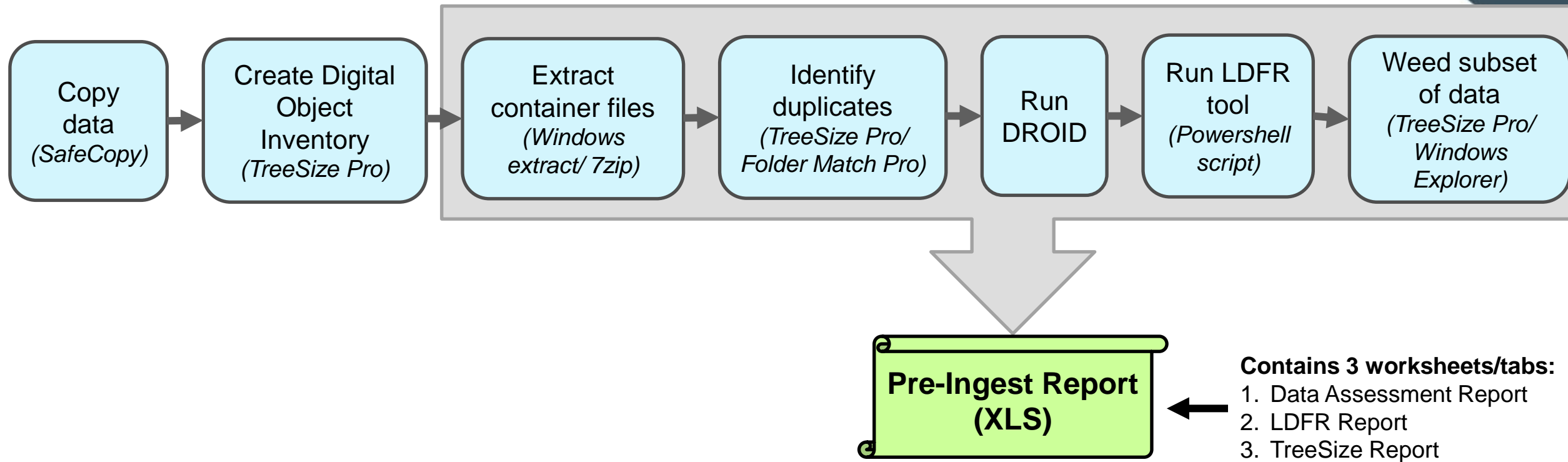
What resources and documentation are available to help me better understand pre-ingest?

Prior to undertaking training on pre-ingest, it is requested that you read the following documents:

1. The Training PowerPoint ([FR / EN](#)) for the Procedures for Pre-Ingest have been added to the slides.
2. The suite of documents related to the [Procedures for Pre-Ingest](#)

# <input checked="" type="checkbox"/>	Task	Notes
3.1 <input type="checkbox"/>	Complete Digital Object Inventory spreadsheet	<ul style="list-style-type: none">• Use TreeSize Pro to create a digital object listing of all the content in the repository. See TreeSize Pro Help Sheet (FR / EN) for directions on how to use the software.• Either reformat the TreeSize Pro export to match the Digital Object Inventory spreadsheet or copy and paste the output from TreeSize Pro into the spreadsheet.• Name or rename the Digital Object Inventory after the Repository Name (e.g. Inventory_2016-5468.xlsx)
3.2 <input type="checkbox"/>	Run DROID on entire registration	<ul style="list-style-type: none">• Create a DROID export (*.CSV) for all of the content in the repository.• Note – if the volume of content transferred is very large (e.g. > 1TB) it will take a very long time to complete<ul style="list-style-type: none">○ In order to manage this, DROID can be run on one or more smaller subsets of the data resulting in multiple CSV files.○ Alternatively, if you are confident in the assessment of the content (e.g. no file format extension) – a judgmental call may be made to run DROID on the entire content identified by TreeSize Pro (via a copy of that content).• See DROID Help Sheet for info on how to use the software.

How do we do Pre-Ingest?



How do we do Pre-Ingest?

Main categories	Description	Software used	Outputs
1. Review of Pre-ingest request	<ul style="list-style-type: none"> Check for missing work/info which is required prior to starting Pre-Ingest <i>e.g. Inventory of physical carriers in transfer, creation of a processing workspace to copy the data</i> 	N/A	Email if additional work is needed
2. Data copying	<ul style="list-style-type: none"> Anti-virus Checksums (MD5) Maintain time & date stamps when possible Troubleshoot issues as needed 	<ul style="list-style-type: none"> SafeCopy TeraCopy MD5summer Fsum Frontend Mac Checksum 	<ul style="list-style-type: none"> Copy log metadata (CSV) Screenshot of virus warnings if present (JPG) Physical Carrier Inventory (XLS)
3. Pre-Ingest analysis (aka "Technical Appraisal")	<ul style="list-style-type: none"> Identify file formats Categorize formats based on format & current LAC capacity Weed subset of content (i.e. system/application files – content donor or government institution did not create or interact with & which is not critical to the rendering/functionality of digital archival records) 	<ul style="list-style-type: none"> DROID LAC's LDFR tool TreeSize Pro 7zip / Windows extract QuickView Plus Libre Office Passware Analyzer 	<ul style="list-style-type: none"> Duplicate listing (XLS) Password & encrypted file listing (CSV) DROID report (CSV) LDFR tool results (TXT) Pre-Ingest Report (XLS)
4. Communicate with archival clients	<ul style="list-style-type: none"> Brief archival staff on Pre-Ingest results (email/meeting) Highlight next steps & any pertinent issues to consider for archival processing Manage expectations from a digital preservation POV 		<ul style="list-style-type: none"> Email messages MS Teams meetings

Data copying - SafeCopy

The image displays the Pinpoint SafeCopy 3.0.675 Desktop Edition interface. The main window shows a toolbar with icons for 'Add Source', 'Add Files', 'Add List', 'Pick Target', 'Resume', 'Copy Options', and 'Help'. The 'Copy Options' icon is circled in red. Below the toolbar is a 'Data Sources' section with a large empty box. At the bottom, there are fields for 'Job Target Path' and 'Log File Path', both circled in red. Arrows point from these fields to the 'Job target path' and 'Log file path' labels. A 'Digital processing workspace' box points to a folder structure in the background. To the right, a 'Progress' window shows 'Processed 19 of 74 files (92 of 439 MB)' and a list of files. A 'Completion' dialog box in the foreground states 'The process has completed.' with statistics: 'Files copied: 74 of 74', 'Bytes copied: 459916026 of 459916026', 'Errors: 0', 'Time Elapsed: 00:00:00:20', and 'Average Speed: 77.1 GB/hr'.

Job target path – Where data is copied to.

Log file path – Where metadata is copied to.

Digital processing workspace

Pinpoint SafeCopy 3.0.675 Desktop Edition - Progress

Processed 19 of 74 files (92 of 439 MB)

Files Copied: 19
GB/hr: 80.84

0:00:15
Committee Meeting Sep28.2011 - Meeting Binder.zip
h\Desktop\Test-Zip2\ORG\BoD\Audit Committee\Sep 28. 2011 Winnipeg MB
copy thread on 0

nder - Audit Committee Meeting Sep28.2011.zip
Nov 24 Teleconference Binder.zip
ust 9 2011 Teleconference.zip
meeting binder Nov 16. 2011.zip

Pinpoint SafeCopy 3.0.675 Desktop Edition

The process has completed.

Files copied: 74 of 74
Bytes copied: 459916026 of 459916026
Errors: 0

Time Elapsed: 00:00:00:20
Average Speed: 77.1 GB/hr

OK

Technical Appraisal – Container files & TreeSize Pro

Name	Size	% of Parent ...	Files	Type
> Office Files a...	13.7 GB	51.3 %	40,469	Documents and files ...
> Mail Files	18.2 GB	35.8 %	28,274	Email messages and f...
> Graphic Files	15.3 GB	7.4 %	5,833	Files containing pictu...
> Internet Files	22.5 MB	1.6 %	1,258	Files related to the ...
▼ Container Files	6.8 GB	1.1 %	897	Compressed Archive...
.zip	6.2 GB	1.1 %	891	Compressed (zipped)
.dmg	160.1 MB	0.0 %	4	DMG File
.z	1.3 KB	0.0 %	1	Z File
.iso	507.9 MB	0.0 %	1	Disc Image File
> Miscellaneou...	1.4 GB	1.0 %	785	Unknown file types
> Program Files	3.6 GB	0.5 %	416	Program Files, Librari...
> Video Files	15.5 GB	0.3 %	224	Files containing vide...
> Text Files	5.2 MB	0.2 %	130	Plain text files, log fil...
> Database Files	430.9 MB	0.1 %	118	Files containing the d...

ADMIN-GOVERNANCE - Finance - BackUp > 2015-2016

Name	Date modified	Type
RYa05392	2016-08-30 2:36 PM	File
Q8a03060	2016-07-26 2:32 PM	File
PLS1223.ZIP	2015-12-23 4:33 PM	Compressed (zip)
PLS1208.ZIP	2015-12-08 2:26 PM	Compressed (zip)
PLS1124.ZIP	2015-11-24 8:47 PM	Compressed (zip)
PLS1105.ZIP	2015-11-05 1:48 PM	Compressed (zip)

Finance > Backup

Name
PLS0805_ZIP
PLS0917_ZIP
PLS1007_ZIP
PLS0129.ZIP
PLS0211_ZIP

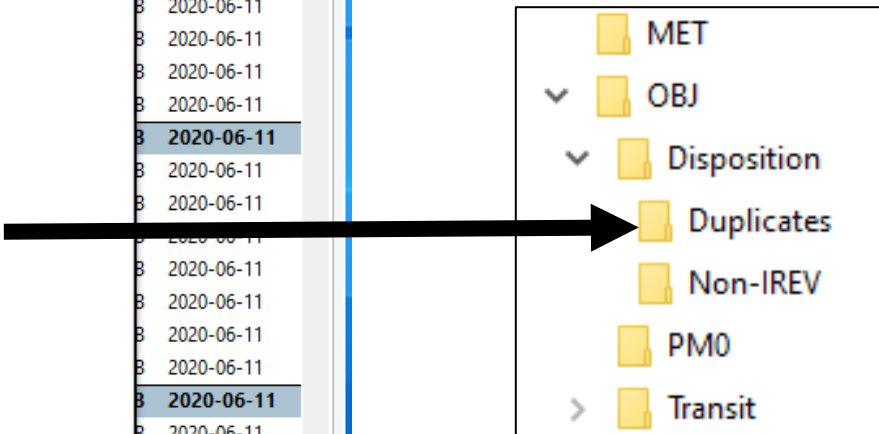
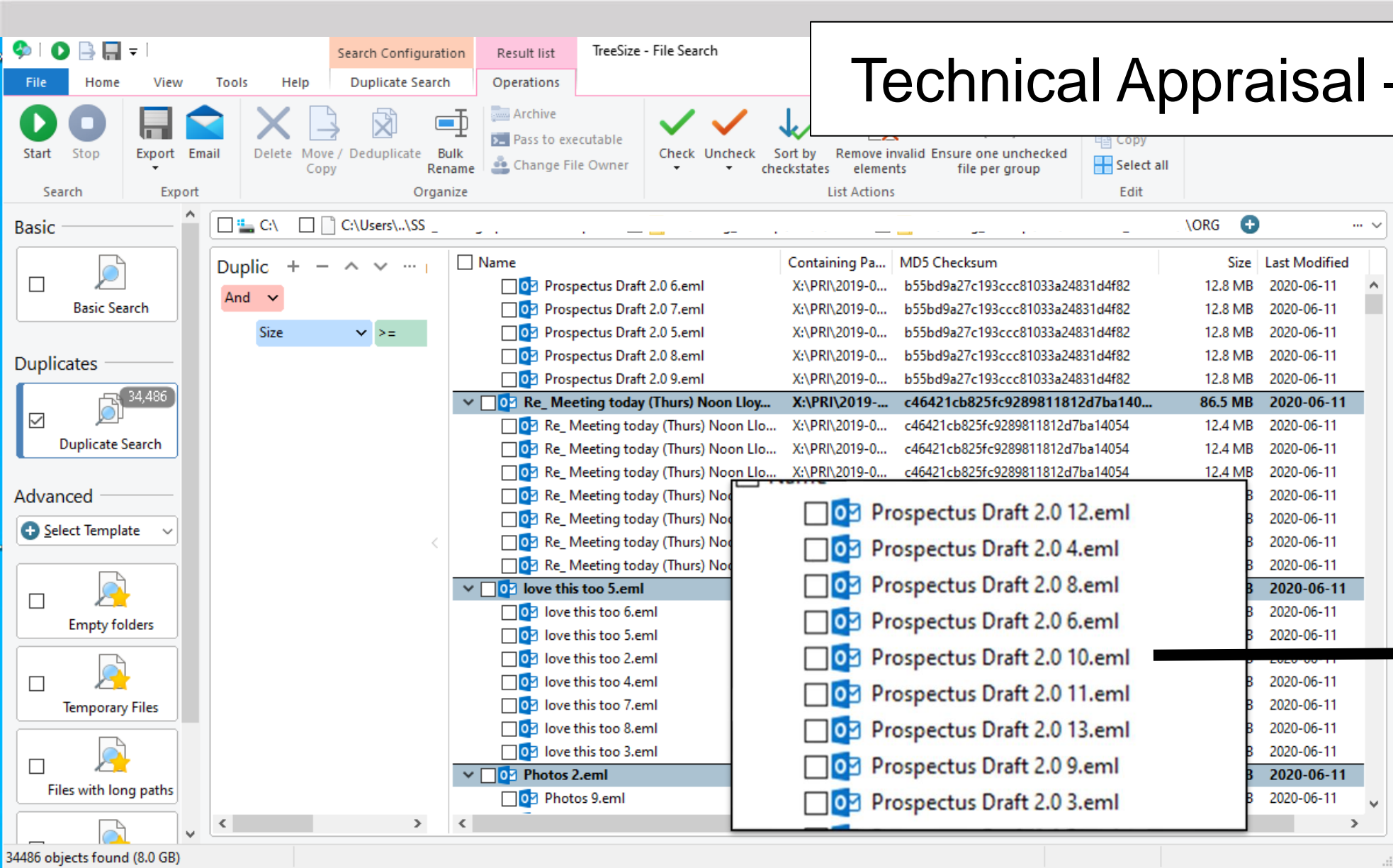
Disposition > Duplicates

Name
PLS0805.ZIP
PLS0917.ZIP
PLS1007.ZIP

Technical Appraisal – Duplicates



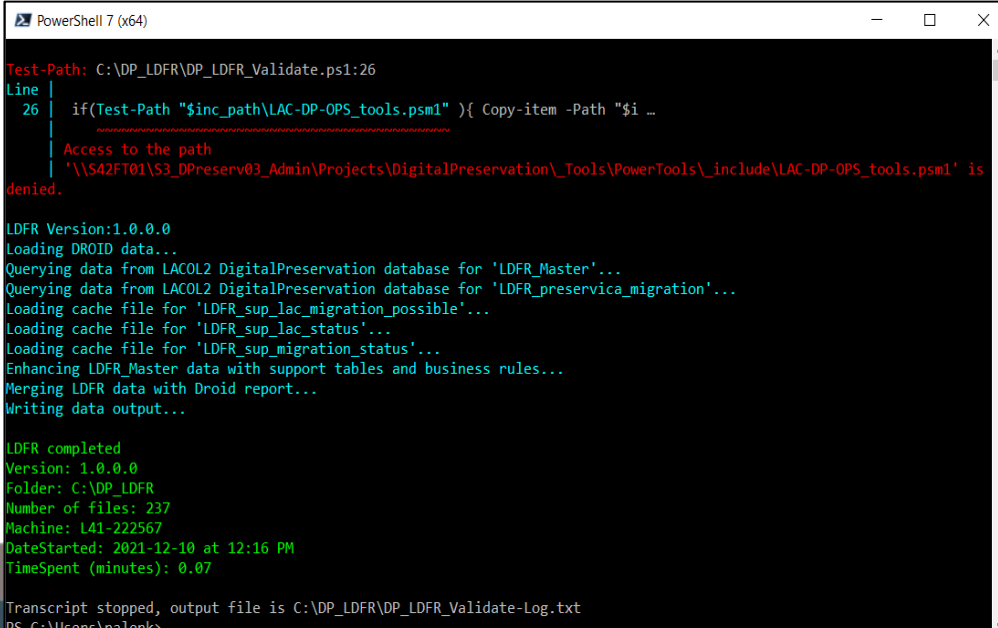
Duplicates search
using checksums &
TreeSize Pro and/or
FolderMatch Pro



Technical Appraisal – File Formats

Use DROID to identify file formats & results exported to CSV

- Past approach → Manual review of formats
- Current approach (2022) → Automated analysis with LAC created LDFR tool
 - LDFR = Local Digital Format Registry
 - Developed by LAC's Digital Preservation section



```
PowerShell 7 (x64)

Test-Path: C:\DP_LDFR\DP_LDFR_Validate.ps1:26
Line
26 | if(Test-Path "$inc_path\LAC-DP-OPS_tools.psm1") { Copy-item -Path "$i ...
    | ~~~~~
    | Access to the path
    | '\\S42FT01\S3_DPreserv03_Admin\Projects\DigitalPreservation\Tools\PowerTools\include\LAC-DP-OPS_tools.psm1' is
denied.

LDFR Version:1.0.0.0
Loading DROID data...
Querying data from LACOL2 DigitalPreservation database for 'LDFR_Master'...
Querying data from LACOL2 DigitalPreservation database for 'LDFR_preservica_migration'...
Loading cache file for 'LDFR_sup_lac_migration_possible'...
Loading cache file for 'LDFR_sup_lac_status'...
Loading cache file for 'LDFR_sup_migration_status'...
Enhancing LDFR_Master data with support tables and business rules...
Merging LDFR data with Droid report...
Writing data output...

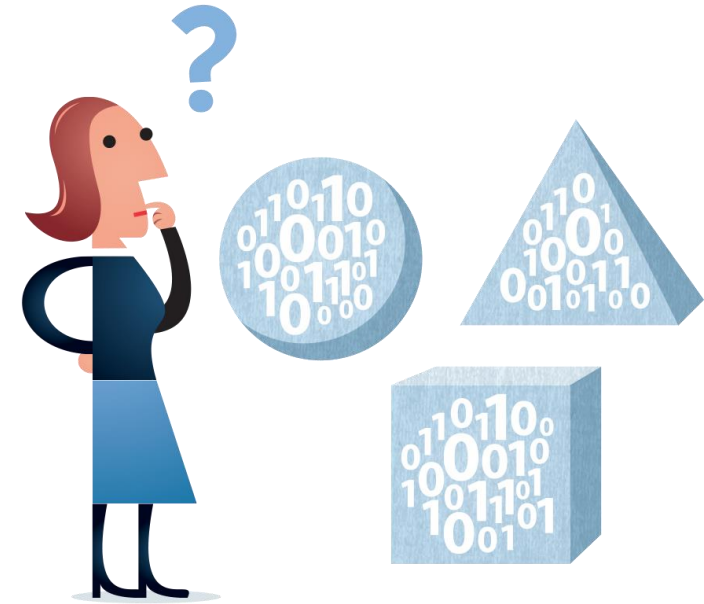
LDFR completed
Version: 1.0.0.0
Folder: C:\DP_LDFR
Number of files: 237
Machine: L41-222567
DateStarted: 2021-12-10 at 12:16 PM
TimeSpent (minutes): 0.07

Transcript stopped, output file is C:\DP_LDFR\DP_LDFR_Validate-Log.txt
PS C:\Users\palenk>
```

Technical Appraisal –File Formats & LDFR Tool

LAC's LDFR tool is a registry of file formats. It includes:

- **LAC's current policy decision on file formats:** e.g. preferred, acceptable, tolerable, not eligible
- **LAC's current capability to manage** said formats



Digitalbevaring.dk

What are the LDFR categories?

Six main categories:

Category	Description
Preferred	No preservation action needed – file format ok as-is
Acceptable	File format can be auto-migrated
Tolerable	File format can be migrated manually
Assessment required	File format has PUID but is new to LAC – research required to determine LAC policy
Cannot assess	File format does not have a PUID or can only be identified by extension – requires digital object level analysis (i.e. 1 by 1 research)
Not eligible	File format is not eligible for preservation

LDFR tool results

LDFR - summary

DIM	File Count	Size (GB)	Distinct Format(s)	Count
A-PREFERRED	74	0.1		5
B-ACCEPTABLE AUTO MIGRATION	122,546	229.9		25
C-TOLERABLE MANUAL MIGRATION	9	0.0		5
D-NOT ELIGIBLE FOR DP	89	0.0		6
X-ASSESSMENT REQUIRED	2,237	0.0		18
Z-CANNOT ASSESS	29,611	47.5		13

LDFR - details

DIM	File Count	Size (GB)	Distinct Format(s)	Count
A-PREFERRED	74	0.1		5
- A0-No Action Needed	74	0.1		5
B-ACCEPTABLE AUTO MIGRATION	122,546	229.9		25
- B1-ACCEPTABLE / Tested Automated Migration	122,546	229.9		25
C-TOLERABLE MANUAL MIGRATION	9	0.0		5
- C1-TOLERABLE / Requires DP staff to Migrate	9	0.0		5
D-NOT ELIGIBLE FOR DP	89	0.0		6
- D1-Non-Archival content files (detected by fileName)	8	0.0		1
- D4-Zero Bytes Files	44	0.0		3
- D5-Else	37	0.0		2
X-ASSESSMENT REQUIRED	2,237	0.0		18
- X1-HAS PUID / UNKNOWN requires analysis	2,215	0.0		15
- X2-Has PUID - Missing in LDFR	22	0.0		3
Z-CANNOT ASSESS	29,611	47.5		13
- Z0-NO PUID Unknown File Format	29,181	47.5		1
- Z1-failed characterization (not signature or container)	430	0.0		12

Pre-Ingest Report

- 1 - Formulaire d'inventaire de supports p
- 1 - Physical Carrier Inventory.xlsx
- 2 - Digital Object Inventory.xlsx
- 2 - Formulaire d'inventaire d'objets num
- 3 - Pre-Ingest Report.xlsx
- 3 - Rapport sur le processus de la pré-ing
- 4 - Quality Assurance Report.xlsx
- 4 - Rapport sur l'assurance de la qualité c

Contains 3 worksheets/tabs:

1. Data Assessment Report
2. LDFR Report
3. TreeSize (content categories) Report

DISCOVER. UNDERSTAND. CONNECT.

Data Assessment Report (Raw data is in LDFR Report tab)

This report assesses the data quality of transferred content. It is based on DROID software and the PRONOM database (an authoritative tool for file format identification).

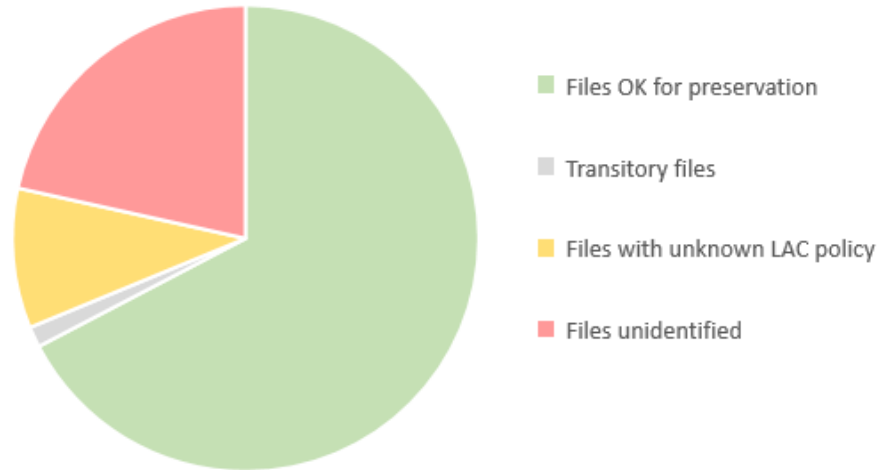
The results categorize the data into 1 of 4 categories depending on LAC file format policy and LAC's current ability to render and/or migrate the data.

Registration #: 2020-XXXX Source: Private
Assessment by: Heather Tompkins Donor/Department: Marc Richard
Archivist Contact: P. Jones Data Location: [\\s42cvma06\FS03-Projects_Prelng\PRI\2020-XXXX_Richard](#)
Date of transfer to LAC: 2015-08-12
Date of assesment: 2022-09-06

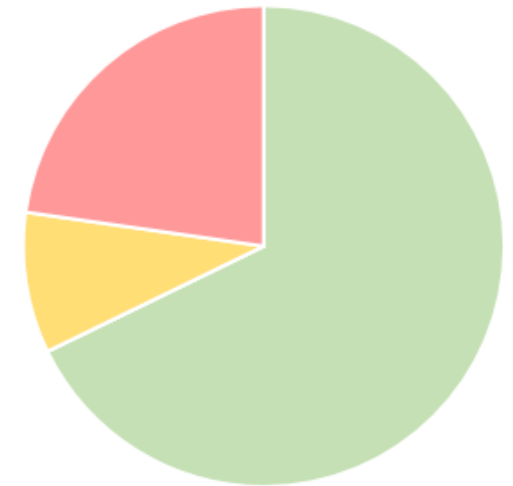
Section 1. Results	Before Pre-Ingest Processing		After Pre-Ingest Processing	
Summary	Number of files	% of payload	Number of files	% of payload
Files OK for preservation	2,361	67.3%	2,221	67.8%
Transitory files	50	1.4%	0	0.0%
Files with unknown LAC policy	340	9.7%	310	9.5%
Files unidentified	755	21.5%	745	22.7%
Totals	3,506	100%	3,276	100%

	General Pre-Ingest Statistic Comparison	
	Before	After
Number of files	3506	3276
Number of folders	257	231
Total size	10.1 GB	8.7 GB

Prior to Pre-Ingest Review



Post Pre-Ingest Review



Definitions:

File format is OK for preservation = Preferred, acceptable or tolerable formats per LAC policy.

Transitory files = System generated files; application/software files; temporary files; 0 byte files etc. Ineligible for preservation per LAC policy

Files with unknown policy = Non-standard file formats not accounted for in LAC preservation policy. Analysis required to determine LAC policy.

Files unidentified = Cannot identify file format. Requires file-level assessment to determine preservation eligibility

Notes:

Data Assessment Report

LDFR Report

Content Categories Report

List Data



LDFR tool → LDFR Report

Files OK for preservation = Preferred, acceptable or tolerable file formats per LAC policy					
Transitory files for removal = System generated files. Ineligible for preservation per LAC policy					
Files with unknown policy = Non-standard file formats not accounted for in LAC preservation policy. Analysis required to determine eligibility.					
Files unidentified = Cannot identify file format. Requires file-level assessment to determine preservation eligibility					
DI Comments	LDFR_status	LDFR_FilePath	LDFR_FileName	LDFR_PU	LDFR_Forma
Content looks to be related to an application. Moved to Disposition\Non-IREV	X2-Has PUID - Missing in LDFR	X:\PRI\2020-XXXX_Richard\OBJ\Transit\ORG\039_flop	PRINT.TST	fmt/1717	Time Stamp
Content looks to be related to an application. Moved to Disposition\Non-IREV	Z0-NO PUID Unknown File Format	X:\PRI\2020-XXXX_Richard\OBJ\Transit\ORG\039_flop	WINSTALL.COM		
Content looks to be related to an application. Moved to Disposition\Non-IREV	Z0-NO PUID Unknown File Format	X:\PRI\2020-XXXX_Richard\OBJ\Transit\ORG\039_flop	WINSTALL.OVR		
Content looks to be related to an application. Moved to Disposition\Non-IREV	Z0-NO PUID Unknown File Format	X:\PRI\2020-XXXX_Richard\OBJ\Transit\ORG\039_flop	WS.COM		
Content looks to be related to an application. Moved to Disposition\Non-IREV	Z0-NO PUID Unknown File Format	X:\PRI\2020-XXXX_Richard\OBJ\Transit\ORG\039_flop	WS.INS		
Content looks to be related to an application. Moved to Disposition\Non-IREV	Z0-NO PUID Unknown File Format	X:\PRI\2020-XXXX_Richard\OBJ\Transit\ORG\039_flop	WSBR.COM		
Content looks to be related to an application. Moved to Disposition\Non-IREV	Z0-NO PUID Unknown File Format	X:\PRI\2020-XXXX_Richard\OBJ\Transit\ORG\039_flop	WSMSG.S.OVR		
Content looks to be related to an application. Moved to Disposition\Non-IREV	Z0-NO PUID Unknown File Format	X:\PRI\2020-XXXX_Richard\OBJ\Transit\ORG\039_flop	WSOVL1.OVR		
	B1-ACCEPTABLE / Tested Automated Migration	X:\PRI\2020-XXXX_Richard\OBJ\Transit\ORG\41\JDD#3	JDD#3-10.896	x-fmt/394	WordPerfec
	B1-ACCEPTABLE / Tested Automated Migration	X:\PRI\2020-XXXX_Richard\OBJ\Transit\ORG\41\JDD#3	JDDRAFT.3-J	x-fmt/394	WordPerfec
System/application file - moved to Disposition\Non-IREV	D4-Zero Bytes Files	X:\PRI\2020-XXXX_Richard\OBJ\Transit\ORG\41\DESK	DESKTOP		
System/application file - moved to Disposition\Non-IREV	X2-Has PUID - Missing in LDFR	X:\PRI\2020-XXXX_Richard\OBJ\Transit\ORG\41\FIND	FINDER.DAT	fmt/1730	Data File
System/application file - moved to Disposition\Non-IREV	Z0-NO PUID Unknown File Format	X:\PRI\2020-XXXX_Richard\OBJ\Transit\ORG\41\RESO	DESKTOP		

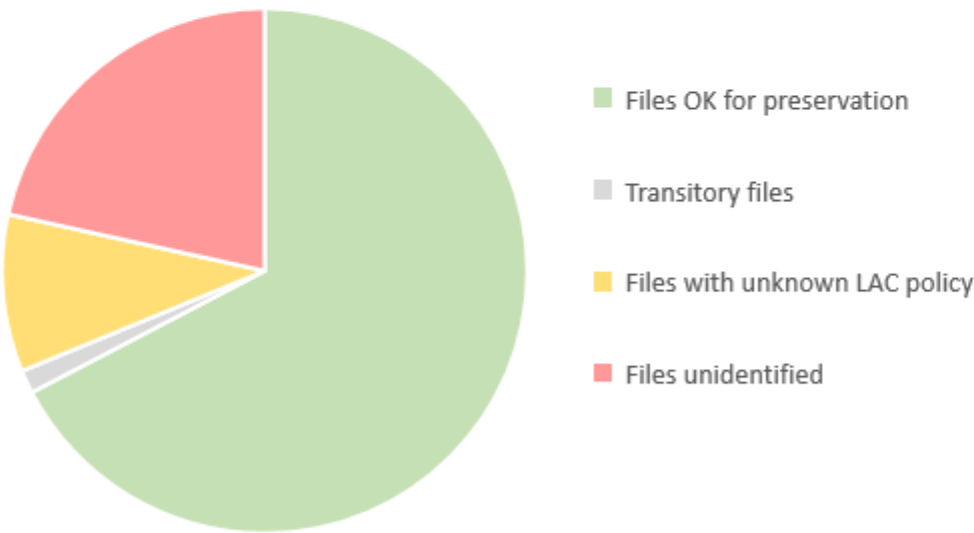
Data Assessment Report (Raw data is in LDFR Report tab)

This report assesses the data quality of transferred content. It is based on DROID software and the PRONOM database (an authoritative tool for file format identification).
The results categorize the data into 1 of 4 categories depending on LAC file format policy and LAC's current ability to render and/or migrate the data.

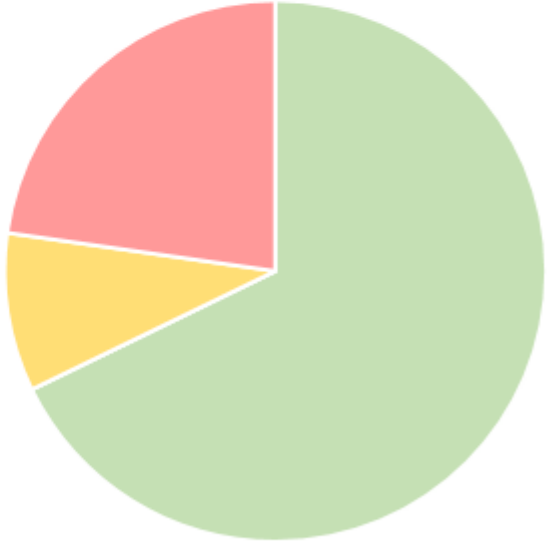
Registration #:	2020-XXXX	Source:	Private
Assessment by:	Heather Tompkins	Donor/Department:	Marc Richard
Archivist Contact:	P. Jones	Data Location:	\\s42cvma06\FS03-Projects_PreIng\PRI\2020-XXXX_Richard
Date of transfer to LAC:	2015-08-12		
Date of assesment:	2022-09-06		

Section 1. Results		Before Pre-Ingest Processing		After Pre-Ingest Processing		General Pre-Ingest Statistic Comparison	
Summary		Number of files	% of payload	Number of files	% of payload	Before	After
Files OK for preservation		2,361	67.3%	2,221	67.8%	3506	3276
Transitory files		50	1.4%	0	0.0%	257	231
Files with unknown LAC policy		340	9.7%	310	9.5%	10.1 GB	8.7 GB
Files unidentified		755	21.5%	745	22.7%		
Totals		3,506	100%	3,276	100%		

Prior to Pre-Ingest Review



Post Pre-Ingest Review



Definitions:
File format is OK for preservation = Preferred, acceptable or tolerable formats per LAC policy.
Transitory files = System generated files; application/software files; temporary files; 0 byte files etc. Ineligible for preservation per LAC policy
Files with unknown policy = Non-standard file formats not accounted for in LAC preservation policy. Analysis required to determine LAC policy.

Transform LDFR results into easy to understand Data Assessment Report

Can inform:

- Additional investment
- Management of resourcing
- Managing expectations re: access

TreeSize Report						Date:		2022-09-02				PRE-INGEST		POST PRE-INGEST			
Registration #:		2020-0132				Source:		PRI				Number of files:		235		184	
Pre-Ingest by:		Heather Tompkins				Donor/Department:		M. Richard				Number of folders:		37		30	
Archivist contact:		John Smith				Pre-Ingest server:		\\s42cvma06\FS03-Projects PreIng\PRI\2020-				Total size:		9.2 MB		7.9 MB	
Category	Result		# of files	Total size	Unit		Action Req'd	Notes							Reference (i.e. hyperlink to spreadsheet)		
Audio files	Yes	▼	64	1.1	MB	▼											
Video files	No	▼				▼											
Container files	No	▼				▼											
Duplicates	Yes	▼	12	956.9	KB	▼	Yes	TreeSize can be use to ID duplicate files. Archivist to determine what to retain.									
Websites	Yes	▼	2	115.1	KB	▼	No	HTM files - should be accessible via a web browser									
Email	Yes	▼	1	207	KB	▼	Yes	Possibly 1 eudora email file (maybe a mailbox?) on carrier 9. If archivist wishes to review/access, contact DI for further work.							\\s42cvma06\FS03-Projects PreIng\PRI\2020-XXXX Richard\OBJ\Transit\ORG\009		
Data	No	▼				▼											
Databases	No	▼				▼											
Password protected files	No	▼				▼											
Encrypted files	No	▼				▼											
0 byte files	Yes	▼	1			▼	No	In Carrier 41 - weeded							\\s42cvma06\FS03-Projects PreIng\PRI\2020-XXXX Richard\OBJ\Disposition\Non-IREV\System-App-Files		
Empty folders	No	▼				▼											
Long filepaths	No																
Non-IREV																	
Deleted files - folder level weeding (\$RECYCLE.BIN; .Trashes; TrueDelete)		▼				▼											
MAC OS - folder level weeding (.fsevents; .Spotlight-V100)		▼				▼											
Temporary files	No	▼				▼											
System/Application files	Yes	▼	50	0.98	MB	▼	No	Files weeded to Disposition\Non-IREV							\\s42cvma06\FS03-Projects PreIng\PRI\2020-XXXX Richard\OBJ\Disposition\Non-IREV\System-App-Files		
Configuration files	No	▼				▼											
Software development files	No	▼				▼											
Thumbs.db	No	▼				▼											
Virus/malware/spyware	No	▼				▼											

Communicating Pre-Ingest Results

- Documentation generated
 - Digital Object Inventory
 - Updated Physical Carrier Inventory
 - Duplicate Listing
 - Pre-Ingest Report
- Email or in-person meeting
- Archival processing commences

Thanks to **Maxime Champagne** (Team Lead, Digital Preservation section) for the development of the LDFR tool!

Further reading on the LDFR tool:

Smyth, Tom. 2022 *Do we really know our data? Assessing file format policy compliance and digital preservation tenability via a new software tool.* IPRES 2022 Conference Proceedings.

Contact info:

Heather Tompkins
heather.tompkins@lac-bac.gc.ca

Our website: www.library-archives.canada.ca

Search our collection: www.collectionscanada.gc.ca

National
Capital Region

Vancouver
British Columbia

Winnipeg
Manitoba

Halifax
Nova Scotia



Library and Archives
Canada

Bibliothèque et Archives
Canada

Canada