Do We Really Know Our Data?

Toward a Methodology for Assessing Preservation Sustainability and Improving PAIMAS Negotiations via Characterization

|  |  |  |
| --- | --- | --- |
|  | **Tom J. Smyth** |  |
|  | *Library and Archives Canada**tom.smyth@bac-lac.gc.ca*[*ORCID 0002-3829-2980*](https://orcid.org/0000-0002-3829-2980) |  |
|  |  |  |

**Abstract – Library and Archives Canada’s (LAC) new File Formats Policy, which constitutes a from-scratch analysis of the Preservation area’s migration capacity, has been converted into a database that can be remotely queried. In tandem with DROID, we developed a new software tool that compares a target dataset (e.g., material for acquisition, or material being migrated to a new media) against our PRONOM-based Local Digital Format Registry (LDFR) database. The tool then outputs a report – which then speaks to LAC’s capacity to manage the target data – and flags detectable file format and data management issues. The tool is a major development for LAC and enables us to assess file format compliance, preservation tenability, and therefore the estimated cost of ownership at the pre-transfer stage of the lifecycle – which will inform PAIMAS-oriented discussions with internal and external clients – leading to better overall capacity management, and digital preservation migration, planning, and sustainability.**

**Keywords – File format policy, characterization and migration, preservation tenability and sustainability, Digital preservation planning and capacity analysis, PAIMAS, PRONOM.**

**Conference Topics – Innovation; Resilience.**

# Introduction

In 2021-22, the Digital Preservation (DP) area of Library and Archives Canada (LAC) undertook a major revision of its File Formats Policy. An institution’s file formats policy is typically a statement of its ability and capacity to manage and preserve data, organized based on the ideal file or data formats it wants to receive for particular content types. Some formats are easy for the organization to manage, while others are not for various reasons. By definition, this is an individualistic policy piece based on the details of an organization’s mandate, perhaps its legal context, its dedicated resources, and to a degree, its DP program maturity. It also tends to be a snapshot, a moment in time of when it was written or published – which can be problematic, as such a core policy of the DP Archive evolves quickly under the lens of operations, so should be communicated to stakeholders at equal pace where possible (we can improve on that ourselves!)

This piece falls into the ‘Innovation’ conference topic insofar as it explores how we analyzed migration capacity and began assessing digital preservation tenability and sustainability at the point of pre-transfer. It also falls under ‘Resilience’ in terms of the strategy we built to characterize capacity challenges and build objective data on which to describe quality and the costs of ownership – while enabling an institutional digital preservation services catalogue, and delivering better and more strategic digital preservation services to our clients.

# Problem Statement

While we have maintained a paper-based policy since at least 2009, we knew we were missing a means of applying it in a machine-readable and dynamic manner to everyday digital operations. Thus, our goal was to construct a method to assess any data under scrutiny for its ‘compliance’ with the File Format Policy – in the contexts of incoming acquisitions data, legacy media and collections under digital preservation (DP) migration, and perhaps also for collections and backlog re-appraisal.

In so doing, we wanted to address two main issues: 1. Creating a means of describing and mitigating digital preservation risk inherent to poor data quality or unmanageable file formats before they came in the door, and 2. Building the data on which to do DP planning, risk assessment, and migration for the DP Archive. Ultimately, we wanted to begin the process of articulating and estimating the costs of a work package or the ownership of any target data – which can be more than financial.

Many great tools exist for file format characterization (DROID [1], Brunnhilde [2], Jhove [3], etc). However as a next step in digital preservation migration program maturity, we found ourselves asking how an organization 1. Analyzes and understands its total and current-state capacity to migrate file formats (in terms of expertise, tools, and resources), 2. How it compares its preferably evergreen file formats policy against .e.g., incoming data for potential acquisition in an automated and machine-readable manner, and 3**.** How this process could assess and describe target data compliance against the File Formats Policy, both in terms of its preservation tenability and the resources it is likely to absorb -- initially and in the long-term.

Put another way, *how do we estimate the overall sustainability of a given acquisition, its risks and its costs of ownership – at the point of pre-transfer, with an eye to the core principles of OAIS sustainability – before an acquisition decision needs to be made? Where we know we must acquire a given collection, how can we articulate the cost of ownership in its current state, and calculate what resources and capacity will be necessary for its management? How should that information impact PAIMAS-oriented client negotiations, where possible?*

Going further, in financial, priority, and capacity management contexts, how do we build a framework that ensures we are applying our finite resources to the digital documentary heritage content that warrants high investment? How do we gather objective data with which to brief stakeholders on what the costs of ownership are likely to be? (and then use it to augment acquisition decisions and/or build business cases for more capacity?) For us at LAC, this was also the beginning of a vision for an institutional DP services catalogue, where the services could be added or removed based on available funds, capacity, and priority.

We therefore produced a methodology and tool that would generate real-time data, and enable the relevant conversations with easy-to-understand reports.

# Analyzing File Format Capacity

An issue we identified over the years was how quickly our document-based File Formats Policy became outdated, and how this affected ongoing acquisition operations since clients were constantly referring to and acting upon the outdated web-based copy. We needed a new means of ‘keeping it evergreen’, and we also wanted to make it capable of specific query (so we were not referring to a paper document on a digital matter).

To address this, and to ensure the data could be easily updated in operations contexts, we generated a relational object and populated it with formats and Persistent Unique Identifiers (PUIDs) based on the UK PRONOM database [4]. We also ensured the database was capable of exporting its information in an eye-readable HTML format. I have the habit of referring to this as the LAC LDFR – Local Digital Format Registry, which also tends to encompass the paper policy document.



LAC utilizes Preservica [5] as its core digital preservation module in its suite of systems and services that we refer to as the “Digital Assessment Management System” (DAMS). A next step in analyzing our institutional capacity for migration was to compile the list of file formats that were preferred for acquisition as-is (i.e., are best for LAC for their content type in the present state of 2022, based on our experiences and in comparison with the file format policies of other organizations). We then added formats to the LDFR database for which there was an existing, or a Preservica migration pathway (i.e., those formats that are not ideal, but whose migration can be automated). Next, we added formats we knew we could migrate but which were dependent on human intervention (i.e., format migration that had to be conducted manually with a desktop application or otherwise needed DP staff expertise). The previous LDFR policy and staff knowledge informed the compilation of these latter “must be manually migrated” formats.

By process of elimination, this created a pool of remaining file formats from PRONOM or the previous policy that required elevated resources for LAC to manage. This was due to 1. An absence of known or deployed tools to handle the formats (e.g., M365!), 2. Those which had been previously flagged as demonstrating some issues (e.g., geospatial, AutoCAD), 3. Formats we knew existed from PRONOM but which have never entered our workflows (which would trigger DP analysis and policy decisions the first time they appear), or 4. Formats known to us but which are absent from the PRONOM database and thus do not have PUIDs (i.e., formats and notes we should contribute!)

By arranging these file formats into lettered categories within the LDFR database, we can move, edit, or update their status on the fly. The LDFR thus shows at-a-glance (or through remote query) what our internal capacity is to handle any particular version of a file format – since each category reflects an increasing scale of complexity, inherent requirements, and a particular ‘swim lane’ for data management within the DP unit.

# Comparing LDFR to Target Data

Having now defined and established what our capacity is to handle individual formats via LDFR database categorization and file format characterization, we needed a means to compare it against incoming acquisitions data (or data in backlogs, or being migrated off legacy carriers), to assess the target data for risk, digital preservation tenability, likely work package cost, and thereby sustainability.

Using PowerShell, the DP area created a tool driven by DROID reports (“the LDFR tool”). Dropping a DROID report into a particular network directory would execute a script that runs a comparison between the DROID report and the LDFR database – and outputs color-coded results in HTML, which categorizes the target data according to the chart above – which speaks to its degree of compliance with the File Formats Policy. The table can be expanded to show all file formats identified with file counts for each category. Where the analyzed data does not comply, the report articulates what percentage and file count of the target dataset is implicated by category – and therefore what DP actions and work are necessary. This makes the LDFR reports highly granular, objective, and machine-readable (based on PRONOM and our current capacities as reflected in LDFR!), and also easy-to-use in facilitating discussions with e.g., archivists about the data they would like to acquire. Often this triggered decisions in consultation that a high-effort but low-value component of the dataset was not worth acquiring, or conversely that a high-value component warranted extra work. What quickly emerged was the primacy of good collaboration and exchange of expertise in good collections management.

These LDFR categories also begin to depict the relative degree of effort, cost and/or accepted or inherent risk required to manage the tagged formats (category “A” requires no action as the ideal state, whereas “C” cannot be automated and necessitates a DP subject matter expert. “X” and “Z” require elevated effort, as they require DP staff to conduct analysis, and so on).

# Deploying the Tool andAssessing Sustainability

In our initial stages, and for pre-transfer, pre-ingest, the digital archival records backlog, and in pre-acquisition contexts, the LDFR tool is intended for the “Digital Integration” (DI) area (the team within the LAC Digital Preservation Division that is responsible for PAIMAS and the stages of the DP workflow up to and including archival Submission Information Packages (SIPs)). The LDFR tool’s output report therefore shows the DI power user how manageable or non-compliant the payload data is against our current-to-the-moment migration and preservation capacities in the partner DP unit (the area responsible for the Archival and Distribution Packages, AIPs and DIPs).

The LDFR tool will now be used in everyday operations within DI and DP to generate on-demand and objective data and evidence with which to advise internal clients on the digital curation tenability and sustainability of any given target data, but also to begin planning for how our mutual capacity is occupied (leading to business cases for more FTEs!) The LDFR categories will also be associated with different levels of treatment in our institutional digital preservation service catalogue[[1]](#footnote-1), which then speaks to effort and duration estimates to conduct any proposed work – since more complex work, represented by elevated LDFR categories, are linked to increased levels of required DP staff expertise and time vs. machine automation. The color-coded LDFR tool’s output report can therefore also be used to brief *clients on* how *the data will be managed* for processing and preservation at LAC.

Rolling all these details up into a briefing for a given client for a particular or potential collection, while factoring the client’s access or project wishlist, would then be packaged as necessitating a certain level (“Gold”, “Platinum”) of DP treatment and engagement from the DP catalogue. Where the collection is at the pre-transfer stage, this information might influence the acquisition decision, contribute a great deal to the PAIMAS-oriented pre-transfer submission agreement, or influence an improvement in overall data quality – in order to mitigate estimated effort, risk, cost, and duration to acquire the target data and/or apply the desired services. At times, this has also led to the client delaying transfer of the data, to enable the combined resources of LAC and the client to be applied to the unanticipated but essential work. Our vision is for this work to enable formulaically calculated effort, duration, and costs for projects and communication to stakeholders, which in turn also assists in DP capacity and migration planning.

In the DP unit, the LDFR tool will be used in the near-future on all data being migrated to new DP archive media. This will describe exactly what data by PUID is on each of our 9600+ magnetic tapes, leading to risk-managed actions in migration from LTO-3-4-6 to LTO-9 in our Tape Library. We would then capture these details in our DP master database – which enables DP Archive planning and migration for distant years, when it must again be migrated forward.

# Lessons Learned

The DP area will begin including the DROID and LDFR tool reports in future AIP payloads, thereby transferring the file format analysis, capacity context, and file format migration decisions forward into our DP archive master database – as additional context and provenance information, that may elucidate preservation decisions made now for the benefit of our successors in the future.

The File Format Policy will now remain evergreen, since the full content of the LDFR database can be exported, in document format, and on-demand. Thus, we can provide a fresh copy whenever clients and other departments in the Government of Canada request – in addition to issuing it periodically. While legacy data is legacy data and a File Formats Policy will not always be consulted, applying the thinking as far upstream as possible is our goal, to enable Government of Canada clients to conduct self-assessment, leading to early analysis and warning for future federal record transfers.

Given the framework that is now in place, we can also begin building the data necessary to estimate our bandwidth requirements for the use of Preservica’s pathways for performing migration work in the Cloud.

Not all collections were created equal in terms of data quality or sustainability, and so not all of them can objectively warrant high investment or can support advanced access options. Capacity and resources are always at a premium. This is perhaps not an easy thing to socialize, even when the digital preservation practitioner deeply understands and shares the collection specialist’s sense of intellectual responsibility for acquiring important but not always great quality data.

It is essential to impart to internal stakeholders that their participation in PAIMAS and digital curation and preservation is not designed to control what they can acquire. It is to ensure that preservation tenability and sustainability is possible in alignment with core OAIS and PAIMAS principles. Digital advocacy is for securing the needed resources to address the prescribed and required DP work, and to improve pre-transfer and pre-acquisition conditions, leading to greater efficiency, capacity and improved (data) quality of life for the collections. If we can improve the state of data prior to transfers via this characterization methodology for compliance measuring and cost of ownership estimation, and mitigate them through PAIMAS negotiations, we can avoid future (but absolutely inevitable) impact on the institutional reputation, elevated resource and capacity consumption, and an inability to access collections in the future. For these reasons, I thought this work was important to publish if conceptually, as this approach would be easy to re-purpose anywhere. Our LDFR tool is driven by DROID and PRONOM, so it is simple to update, and does not require any adjustments to the rolled-out tool.

Acquiring data we cannot provide access to is against the cardinal rule of digital preservation. All stakeholders should therefore be concerned with data quality. To ensure informational continuity, digital preservation imposes requirements on data – it must after all, if presumes to ensure access, be tenable in the long-term, or at all sustainable.

#### ACKNOWLEDGMENT

The highest praise is due to the warrior women and men comprising the Digital Integration unit at LAC. I remain humbled and privileged to work with you, as has been the case since 2013. Your digital curation and web archival expertise are unmatched, and on day you will be viewed as among LAC’s digital pioneers.

The work underpinning this paper was conducted in tandem with Maxime Champagne and the *berserkers* of the Digital Preservation unit, who do a great deal of heavy lifting on a daily basis. I am honored to work with you.

# REFERENCES

1. The National Archives. *Digital Record Object Identification Software Tool (DROID).* https://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/
2. Canadian Centre for Architecture and Tessa Walsh. *Brunnhilde: Siegfried-based characterization tool for directories and disk images.* https://github.com/tw4l/brunnhilde
3. Harvard University and Open Preservation Foundation. *JHOVE Harvard Object Validation Environment. https://jhove.openpreservation.org/#:~:text=JHOVE%20is%20a%20file%20format,A%20command%2Dline%20interface*
4. The National Archives. *PRONOM.* https://www.nationalarchives.gov.uk/PRONOM/Default.aspx
5. National Digital Stewardship Alliance (NDSA). *Levels of Digital Preservation*. https://ndsa.org/publications/levels-of-digital-preservation/
1. Our vision for a DP services catalogue was inspired by the NDSA Levels of Digital Preservation Treatment. [↑](#footnote-ref-1)