Format Identification In Context: Patterns & Hazards in Digital Preservation Workflows

Andrew N. Jackson
Digital Preservation Coalition
United Kingdom
andrew.jackson@dpconline.org
0000-0001-8168-0797

Paul Wheatley
Preserve Together Ltd
United Kingdom
paul@preservetogether.com
0000-0002-3839-3298

Steve Daly
The National Archives (UK)
United Kingdom
steve.daly@nationalarchives.gov.uk
0009-0002-2803-3800

Heather Tompkins
Library and Archives Canada
Canada
heather.tompkins@bac-lac.gc.ca
0009-0009-7640-6297

Tyler Thorsted
Brigham Young University
USA
thorsted@byu.edu
0000-0003-0292-0962

David Clipsham
Preservica
United Kingdom
david.clipsham@preservica.com
0009-0006-2611-8877

Micky Lindlar
TIB Leibniz Information Center for Science and Technology
Germany
micky.lindlar@tib.eu
0000-0003-3709-5608

Kieron Niven
Archaeology Data Service
United Kingdom
kieron.niven@york.ac.uk
0000-0002-0537-9238

Kieran O'Leary
National Library of Ireland
Ireland
koleary@nli.ie
0009-0009-7485-0634

David A. Russo
British Library
United Kingdom
david.russo@bl.uk
0000-0003-2829-3936

Ross Spencer
Digital Preservation Consultant
Germany
exponentialdecay.digipres@gmail.com
0000-0002-5144-9794

Leontien Talboom
Cambridge University Library
United Kingdom
lkt39@cam.ac.uk
0000-0001-7408-5471

Amanda Tomé
Digital Research Alliance of Canada
Canada
amanda.tome@alliancecan.ca
0009-0007-9542-3143

Krista Oldham
Texas A&M University
United States
koldham@tamu.edu

Stephen Abrams
Harvard University
United States
stephen_abrams@harvard.edu
0000-0003-2326-6672

iPRES 2025: The 21st International Conference on Digital Preservation, Wellington, Aotearoa New Zealand.

iPRES 2025

0009-0005-7143-6527

Kristina Ford
University of Erlangen-Nuremberg
info@kristiford.com
0009-0007-1988-110X

Stefano Allegrezza
University of Macerata
Italy
stefano.allegrezza@unimc.it
0000-0002-7319-2483

Lucy Wales
British Film Institute
National Archive
UK
lucy.wales@bfi.org.uk

Format identification is crucial for digital preservation, yet because of its complexities it is inconsistently implemented in our workflows. This paper shares findings from the "Registries of Good Practice" project and preservation practices collated by the Preservation Registries Special Interest Group. Analysing diverse institutional workflows, we demonstrate how preservation goals, institutional capacities, and specific preservation stages dictated identification requirements. This paper emphasizes the need to connect disparate practices and establish evidence-based improvements for format identification.

Submission Type **– Full Paper.**

Keywords **– Collaboration, Assessment, Identification, Workflows, Risk Identification**

Conference Topics **– Tūhono - Connect.**

I. Introduction

Format identification is a critical step in every digital preservation effort that seeks to maintain usability over time. We can't know if our bitstreams are worth keeping, or how to preserve them effectively, unless we can understand the content they hold. Divining this requires interpreting the bitstreams with appropriate software. The notion of 'format' is what links bitstreams to the software we depend on to unlock the encoded information.

Format identification tools and registries of format information have long been recognised as a crucial part of the practice of digital preservation [1]. In the years since then, many projects have attempted to build new tools and registries, but few outside of The National Archives and PRONOM have left a sustainable legacy. Findings of the "Registries of Good Practice" project show that practices remain inconsistent, and the tools produced by research projects rarely make it to wider adoption.

As part of the "Registries of Good Practice" project[1], a group of individuals and institutions with a mixture of established and emerging digital preservation services have come together to try to better understand our shared needs by analyzing the workflows where format identification takes place. If we grasp how context affects our requirements, we can help make sure the right tools are chosen for the right tasks, and connect individuals and institutions through their shared needs.

This paper documents our findings, exploring how the requirements and behaviour of the format identification process depend on contextual factors like the goals of wider digital preservation workflow and the current point in the lifecycle.

Pre-ingest content analysis [2] is found to be a particularly important stage, where format analysis is a critical operation within the wider context of detecting 'preservation hazards'. These are issues with digital items that need to be considered as early as possible in order for preservation to succeed.

This paper documents preservation hazards that have been identified across all the contributors. In particular, multipart file formats (where a single logical resource is composed of multiple related files) are identified as a critical shared challenge. This example is used to explore how this trivial task varies across contexts, and to help illustrate some of the barriers to advancing the state of the art in this area.

Finally, we review the current landscape of format identification practices and outline the next steps that would help connect together our disparate

[1] https://github.com/digipres/registries-of-practice-project

practices and establish a solid evidence base for future improvements.

II. Workflows

Through the "Registries of Good Practice" project and the Preservation Registries Special Interest Group[2], a variety of individuals and institutions come together to better understand the range of information sources and tools that the practice of digital preservation depends on.

The interest group began with its most commonly used specialist format registry: PRONOM from the The National Archives of the UK (TNA)[3]. But as the discussion widened to include all of the institutions represented by the author list of this paper, new sources of information and tools came into view.

Through these online discussions, combined with deeper dives through in-person meetings and site visits, it became clear that there are two distinct classes of workflows involving format identification. The expectations for format registries and the behaviour of format identification tools vary depending on whether the goal is to better (A) understand the characteristics of a digital collection, or (B) to automate the processing and preservation of a digital collection.

1. *Understanding the Bitstreams*

When an archive receives deposits or acquisitions, or reports on its existing holdings, workflows are required to scan and characterize the digital content. Analysis of the results will then inform human-led processes such as appraisal, risk assessment and prioritization, preservation planning/action, and reporting.

Pre-Ingest Processing

The most common and important case across the participants in this work was pre-ingest processing, especially where institutions are working with depositors to appraise submissions. This is simply because submission time is often the only opportunity that the original producers of the content are available to resolve any issues.

This is the primary use-case that the DROID file profiling tool[4] is designed for, along with the PRONOM format registry that DROID depends on. The majority of contributors used PRONOM in this manner, usually via DROID, but also via Fido[5] or Siegfried[6]. Crucially, this approach assumes that expert assistance is readily available when needed, and therefore any format-related issues discovered during appraisal can be dealt with.

Reporting

The second most common case was on the reporting of existing holdings. Here, the available metadata (usually generated during pre-ingest or ingest) is aggregated and summarized in order to help an organization understand its holdings and prioritize future work. Tooling in this area is quite inconsistent. Most preservation systems do contain all the relevant information, but only some offer detailed views or analysis of it, with others just acting as sources for locally-developed analysis tools.

Furthermore, as tools improve and information sources evolve, relying on the metadata generated at ingest time becomes increasingly untenable. In this case, those who care for the digital content want to be sure they have an up-to-date understanding of their holdings and are aware of any newly discovered issues or risks. However, tooling for this use case is immature and many institutions do not have the ability to rescan their holdings.

2. *Processing the Bitstreams*

The other class of workflows are those that focus on automated data processing. The purpose of the format identification part of the workflow is to immediately route a given bitstream to a specialized software system or module to perform an automated preservation process. This can be at any time in the lifecycle of digital content. It can be at ingest time, as implemented by, for example, Rosetta, Archivematica, Preservica and the BNF's SPAR for the purposes of normalization. It can be as part of a periodic re-analysis of holdings, as implemented by, for example, Preservica's Active Preservation [3], Rosetta's Preservation Management workflows [4] or the UK Web Archive

[2] https://www.dpconline.org/digipres/pr-sig

[3] https://www.nationalarchives.gov.uk/pronom/

[4] https://www.nationalarchives.gov.uk/information-management/manage-information/policy-process/digital-continuity/file-profiling-tool-droid/

[5] https://github.com/openpreserve/fido

[6] https://github.com/richardlehane/siegfried

[5]. Or even at access time, as is common for the re-writing of web archiving resources [6], or when attempting to integrate emulation tools into access workflows [7].

III. Limitations & Consequences

In an ideal format identification scenario, the format of every bitstream would be uniquely identified, unambiguously matching a given resource to the information and versions of software needed to understand it. In this case, the distinction between types of workflows would no longer matter.

But perfect format identification is impossible. Even if we rule out things like deliberately concocted multi-format files (a.k.a. polyglots) [8], our files and formats are interlinked with the expectations of the environment they are created in or intended for.
Even if formats were uniquely identifiable in principle, the ways we have to identify them are imperfect. We cannot run every possible format parser on every single file or manually review each file. To work out what is what, we need to use the available information to estimate what the most likely format or formats are.

3. *How Format Identification Works*

Format identification relies on a combination of external factors and internal signals. The former usually refers to file extensions, but depending on the platform, this can include Macintosh Type and Creator Codes [9], HTTP Content Type headers [10], or filesystem metadata, such as extended attributes or registered UTIs.

In contrast, internal signals are characteristics of the bitstream that are indicative, or ideally diagnostic, of a specific file format. This typically involves matching short 'magic byte' sequences that are found in particular file formats (as used by PRONOM), but can mean any suitable heuristic, all the way up to running full parsers for some set of formats (as JHOVE does).

Sometimes, the two sources of information lead to inconsistent results. This usually shows up as a file extension mismatch, where the external file extension does not appear to match the format. As internal format indicators are closer to how software parses files, they are usually more reliable than external indicators.

But in many contexts, like individual desktop user configurations or source code packages, internal format identifiers are difficult or impossible to devise because of the nature of the formats themselves. However, when these files are plain-text, or organically structured, such as user configuration files, source code, or research data, the case can be made that the file extensions that creators have chosen should be considered authoritative. Worse still, in many cases, practical internal format indicators are difficult or impossible to devise, due to the nature of the formats themselves. In those cases, we have to rely on file extensions alone.

IV. Format Identification In Context

The next step is to understand how the practical realities of format identification affect the outcomes of different types of workflows. To this end, we start by gathering together some specific cases drawn from the experiences of the group.

1. *Conflicted PDFs*

PDFs are widely used and arrive in institutional collections in many different ways. As PRONOM and DROID are optimized for assisting during record transfer, the binary signatures used to detect PDFs are quite strict. In addition to expecting a versioned identifier at the start of the file (`%PDF-1.6`), the signature also expects an end-of-file marker (`%%EOF`) within 1024 bytes of the end of the bitstream (as per the PDF specification).

In the context of records transfer, any PDFs that don't match these basic factors of formal format conformance are highly likely to be a reflection of some problem in the generation process. Any such issues need to be raised and resolved immediately.

But re-using these 'strict' signatures elsewhere can cause issues. For example, when attempting to route bitstreams to suitable software for viewing, we want to know the likelihood that something is a PDF. Failing to identify truncated files or files with extra 'trailer' data causes surprising blockages in processing workflows, and focuses a lot of time and attention on apparent 'problematic' files that usually do not require repair to be perfectly renderable as they are [11].

2. *Ambiguous PHP*

Many formats, especially text-based formats, are not easy to identify using internal binary or container signatures. Although early testing with AI shows promise [12], in these cases, the PRONOM based tools fall back on matching based on file extension only. If more than one format has the same file

extension, then the format identification process will return multiple matches.

This can be particularly challenging as we seek to extend our format registries to add support for more formats. As more are added, the likelihood of extension collisions grows, then files and formats once uniquely identified will suddenly become ambiguous. Worse still, because the most popular formats are most likely to be documented first, these growing registries likely reflect results that were initially accurate, but became misleading with ambiguity.

This case has already happened with PRONOM, when a recent update to the signatures added a new format with the .php extension. This new format record corresponds to a whole family of formats related to a piece of software used by a single print shop company. This extremely rare format result subsequently started showing up in format results on an equal footing with the much more widespread PHP Hypertext Processing language. In that specific case, the print shop file formats have container signatures, so DROID can be modified to use that additional information to rule the records out. But the fundamental issue remains, and such collisions are certain to arise unless other approaches to format identification are used.

3. *Floppy Disk Formats*

When imaging floppy disks, there are two approaches: either the use of a defined disk image, such as .dsk or .img; or, and this is especially relevant for the more obscure floppy disk formats, where flux streams can be used[7]. However, the actual files on these flux streams are not accessible without the correct encoding of the disk image itself (e.g. information on the number of sectors and tracks, the density of the disk, etc.). Currently, no file format identification tool, even for the encoded disk images, has sufficient information for these types of disk images. *.img* being an example of this.

4. *Audiovisual Pre-Ingest*

In regards to old media, we are forced to rely on contextual clues to determine the formats we are dealing with. In specific domains, the context can be relied upon to focus the identification process. Specifically, workflows dealing with audio and video formats tend not to require generic format identification tools.

By keeping incoming material separate based on the source and using agreed standards, A/V archives are able to 'skip' broad format identification and shift focus to deeper analysis:

- The BFI National Archive validating incoming IMP files from streaming platforms to ensure they are complete and compliant with SMPTE specifications [13]
- The BFI National Archive MediaConch and RAWcooked workflows[8]
- The BFI National Archive System for Television Off-air Recording Archiving (STORA)[9], combining custom scripts and various open source tools [14]
- The IFI use of MediaInfo to gather metadata, spot truncated files and sponsored updates to the software to interpret multi-file Digital Cinema Packages as a whole.[10]

Integrating tools would just confuse things. But if such materials are received into a more general collection, they can be difficult to spot, especially the multi-file formats. Furthermore, the need to capture file format information via a unique identifier, such as a PRONOM Unique Identifier grows with the number of formats obtained within a collection.

V. Hazards

Two main classes of hazards were uncovered while comparing workflows across organizations: hazards relating to the content itself, and hazards relating to our understanding of the content.

7 https://wiki.techtangents.net/wiki/Floppy_Disk_Imaging

8 https://www.dpconline.org/events/eventdetail/114/-/

9 https://blog.bfi.org.uk/knowledge-and-collections/start-with-open-source/

10 https://github.com/MediaArea/MediaInfoLib/issues/992

1. *Damaged, malformed or dangerous content*

There are a range of hazards that can occur due to problems creating or transferring content. Some of these are relevant in any context:

- Damaged bitstreams, e.g. files accidentally truncated during transfers, or from damaged or aged media like floppy or optical discs.
- Bitstreams with associated checksums that do not match.
- Malformed information packages, e.g. files missing from inventory manifests.
- Unexpected encryption/DRM, (for example encrypted fonts in unencrypted ePubs).
- Large .ZIPs over 4GB (often malformed, depending on the source, so need to check if they work)[11].
- Viruses and other malware.

While these issues were considered relevant across all contexts, the mitigation for each may still depend on the goals of the institution and the institution's cultural and legislative context. For example, the UK Web Archive quarantines suspected web viruses as historical artefacts rather than removing them from the historical record. Similarly, some institutions view the presence of system files as providing important contextual information for later access efforts.

Other digital content hazards are more context-dependent:

- 'Too small to be right', e.g. very short audio files, very small image files.[12]
- Zero byte files.
- Blank images or videos, silent audio files.
- Invalid bitstreams, i.e. those that don't conform to the expected format or profile for the context.

For example, if you are working with a digitization programme that is creating reasonably uniform outputs, it makes sense to use the expectations for that content to guide the design of additional checks[13].

2. *Unknown, misunderstood, or missing content*

The second source of hazards is related to making sure the preservation nature of the bitstreams is understood. This means knowing what software is required to access the content, and whether there are other more subtle dependencies hidden within the bitstreams. For example, contributors to not understanding the content include:

- Failed format identification
- Weak identification (e.g. extension only matching, multiple ambiguous matches)
- Mismatched internal/external signatures (e.g. file extension)
- Embedded files [15]
- Multipart file formats not recognized, leading to files getting separated or lost
- Missing components needed or referenced, but not included
- Required file system metadata (e.g. file system date as it might be the only date information you have. Macintosh resource forks, or extended attributes)
- Non-standard characters in file or folder names which impedes preservation actions or accessing the content
- Unusual/older text encodings (a.k.a. mojibake)
- Formats that are strange, difficult to access or obsolete
- Formats that render/appear differently on different operating systems or applications.
- System files, especially ones like Thumbs.db that can contain personal information[14].

VI. Multipart File Formats

The discussion that led to the list of hazards above began with a particularly difficult case: multipart file formats. These have long been recognized as an important case, with JHOVE2 development naming them as 'Clumps' and supporting the detection of the Shapefile format. After that, apart from the specific support for DCP

[11] https://www.bitsgalore.org/2020/03/11/does-microsoft-onedrive-export-large-ZIP-files-that-are-corrupt

[12] They may still be valid, as shown by https://github.com/mathiasbynens/small, but images only a few pixels or bytes in size may indicate issues in a digitisation pipeline.

[13] https://github.com/bitsgalore/jprofile

[14] https://groups.google.com/g/australasia-preserves/c/o110geMSECI/m/e2L8BciuBwAJ?pli=1

added to MediaInfo as mentioned above, there has been very little progress in this area.

This is important, because the failure to identify these related files then affects later ingest and access processes. When related files are not tied together at ingest time, they can become 'atomized' by the ingest progress, with each being treated as a separate item in a way that makes reconstituting the original item extremely difficult. The specific example of Apple Application Bundles was given, the details of which are sufficiently complicated to deserve a separate paper [16]

As such, the group began to investigate the range and complexity of the multipart file sets. The original idea was to gather the information on a number of different file sets and start to build an implementation of a file-extension based multipart file set identification system. However, it rapidly became clear that there were far too many examples to survey quickly, and even in mature cases, it was often difficult to find authoritative information about the precise expected file extension arrangement. Therefore, the group's attention switched to attempting to establish a broader understanding of the nature and consequences of the issue, as outlined here.

Based on the examples and information gathered so far, we propose the following groupings of filenames of multipart file formats to aid in recognizing required and optional files that constitute the multipart content:

- **Required:** Here, the consuming software expects multiple files in particular formats. No parsing is required to extract the file names. The relationships can be definitively determined from the folder and filenames and extensions. For example:
    - Shapefile's shp shx dbf files
    - DVD Video's VIDEO_TS.IFO and VIDEO_TS.VOB
    - Sound Designer II files and a corresponding Resource Fork
    - Agilent ChemStation .uv and .ch files
    - The Bag-It directory structure and required files
- **Optional, Implied:** Additional files that may be included alongside a Required file, but are not directly named or referenced in other files. For example:
    - DVD Video BUP (backup versions of IFO files)
    - Files like *.E* in the context of a *.E01 Encase (required) file.
    - Split ZIP, PAR, RAR, etc.
- **Optional, Referenced:** Additional files that may be part of the multipart set alongside a Required file, but where the identity of those files can only be determined by parsing one or more of the Required files. For example:
    - PDF fonts
    - Images/css/javascript referenced in HTML
    - Project files. InDesign, QuarkXPress, Final Cut, etc.

Any given Multipart File Format will have at least two associated filename conventions, at least one of which is Required. As with singular File Formats, there may be additional arbitrary files that may be referenced from the contents of these files.

● CONCLUSION

Traditionally, work on format identification has often focused on whether formats are obsolete, or on migration to 'preservation formats'. However, our work indicates that a much larger concern is present: the failure to understand our collections well enough to capture important contextual information, software or runtime dependencies. We cannot preserve what we never realized we needed.

Furthermore, format identification is not a singular, universal process. The desired behaviour depends on whether the goal is (1) to survey and understand the content, or (2) to automate the processing of items for information extraction or format migration or emulation.

These are very different uses of the same information. PRONOM and DROID's primary use case is (1). Not that (2) is not also important, but as users of Preservica, TNA is aware that it will be updated, so any disparities or delays are not time critical. However, other users of systems based on PRONOM have found problems arise during automated actions due to issues including ambiguous identification.

For automated processing, tools that return the singular, more generic identification results (e.g. it's probably a PDF, version unknown) are easier to work with, such as libmagic/file (as used by the BNF's

SPAR) or Apache Tika (as used in combination with DROID by the UK Web Archive).

The sheer range and complexity of formats also indicates that it often makes sense to rely on contextual information to reduce the range of formats considered. For example, GitHub Linguist and Google's Magik are relatively good at identifying text-based formats, but directly adding them into existing workflows may cause problems.

This complexity also means that the authoritative information about file format may live outside of the file, and only be something that can be determined from the context. As such, it is critically important that all digital preservation systems and services allow the format of a digital resource to be authoritatively defined by the operator, rather than relying on the assumption that full automation is possible.

This workflow dependence has also helped us identify a wide range of potential hazards that digital preservation initiatives should be aware of when attempting to preserve digital content. This can then form the basis of a detailed understanding of the minimum requirements for things like pre-ingest processing workflows and tools.

Across all contexts, however, we have identified that multipart file formats are a particularly pressing hazard that would benefit from further exploration. Relatively simple tools that can spot file naming conventions could bring huge benefits, but pulling together the authoritative reference information to make this possible will require significant investment of time and energy across the digital preservation community.

References

[1] Abrams, S. "The role of format in digital preservation", VINE, Vol. 34 No. 2, pp. 49-55. https://doi.org/10.1108/03055720410530997

[2] Monks-Leeson, E. and Tompkins H. (2025) "Preservation as access: Digital preservation evaluation and the pre-ingest workflow at Library and Archives Canada" Journal of Digital Media Management"

[3] O'Sullivan et al. "A QUESTION OF CHARACTER: How do we automatically recharacterize data at cloud scales?", iPRES2023 Conference, https://hdl.handle.net/2142/121087

[4] "Validation stack processes", TIB wiki, https://www.google.com/url?q=https://wiki.tib.eu/confluence/spaces/lza/pages/63768010/Erhaltungsplanung%2BPreservation%2BManagement&sa=D&source=docs&ust=1744476209241631&usg=AOvVaw0d4HTl9hk2MDFxu-ax_D9p

[5] Jackson, A. "Formats over Time: Exploring UK Web History", iPRES2012, https://phaidra.univie.ac.at/o:293834

[6] Berlin, J et al. "To Re-experience the Web: A Framework for the Transformation and Replay of Archived Web Pages", ACM Transactions on the Web, Volume 17, Issue 4, https://doi.org/10.1145/3589206

[7] Cochrane, E. "Useable Software Forever. The Emulation as a Service Infrastructure (EaaSI) Program of Work" iPRES2022 Conference.

[8] Albertini, A. "This PDF is a JPEG", POC||GTFO 3, https://mcfp.felk.cvut.cz/publicDatasets/pocorgtfo/contents/articles/03-03.pdf

[9] Thorsted, T. 2024. Macintosh Type/Creator Codes: Improving Identification of Files from MacOS Classic. iPRES 2024 Conference .https://doi.org/10.21428/5676bf2d.76aad2e6

[10] W3C. "The Content-Type Header Field", https://www.w3.org/Protocols/rfc1341/4_Content-Type.html

[11] Töwe et al. "To Act or Not to Act – Handling File Format Identification Issues in Practice", iPRES 2016 Conference, https://hdl.handle.net/11353/10.503183

[12] Green et al, "Text File Format Identification: An Application of AI for the Curation of Digital Records" iPRES 2021.https://phaidra.univie.ac.at/o:1424885

[13] Davis et al, ""You oughta be in pictures": Insights to Digital Moving Image Preservation from the BFI, EYE, and LOC". iPRES2024 panel session. https://www.digipres.org/publications/ipres/ipres-2024/papers/you-oughta-be-in-pictures-insights-to-digital-moving-image-prese/

[14] White, J. "System for Television Off-air Recording and Archiving, BFI National Television Archive" FOSDEM 2024, https://archive.fosdem.org/2024/schedule/event/fosdem-2024-2177-system-for-television-off-air-recording-and-archiving-bfi-national-television-archive/

[15] Allison, T. "Embedded Files: Risks, Challenges and Options", https://www.slideshare.net/slideshow/embedded-files-risks-challenges-and-options/253213162

[16] Thorsted, T. "Macintosh Type/Creator Codes: Improving Identification of Files from MacOS Classic." iPRES 2024 conference. https://doi.org/10.21428/5676bf2d.76aad2e6

Author Biographies

There is insufficient space for a biography of all our authors, instead we will refer to the Preservation Registries Special Interest Group which brings together experts from across the digital preservation field (including all of our authors) to discuss this topic.

Acknowledgments

”Registries of Good Practice” is a two-year research project co-funded by Yale University Library and the Digital Preservation Coalition.