EXTENSIVE EXTENSIONS

Exploring File Extensions in Library of Congress Collections

Mark Cooper
Library of Congress
United States of America
mcop@loc.gov

Aly DesRochers
Library of Congress
United States of America
alde@loc.gov

Jesse Johnston
Library of Congress
United States of America
jesjohn@loc.gov

Camille Salas
Library of Congress
United States of America
csala@loc.gov

Trevor Owens
Library of Congress
United States of America
trow@loc.gov

Abstract – Through four decades of digital initiatives and collecting programs the U.S. Library of Congress has built up a sizable digital collection. In support of long-term management of this digital content, in 2018 staff worked to review information about file extensions of content in the permanent digital collection through analysis of data in the institution's primary digital content inventory system. This paper reports the results of this analysis and how these findings will inform the development of digital content management policy and practice at the institution.

Keywords – file extensions, digital inventory management, digital collections

Conference Topics – Designing and Delivering Sustainable Digital Preservation

I. INTRODUCTION

In 2018, Library of Congress staff began working to identify the major file formats that comprise the institution's permanent digital collection. This paper reports on the initial results of that work, which included an analysis of file extension metadata recorded in the Library's primary digital collections inventory system.

This content is spread across multiple storage architectures and inventorying platforms, and the computing time and ongoing work to process new content makes direct analysis of the files unattainable with present tools and systems. Instead we began work by querying file extension information (e.g., .pdf, .mp3).

This paper describes the methods used to conduct this preliminary analysis and the results. We begin by briefly contextualizing this work within the Library's considerable history of digital initiatives, its digital collecting programs, and the development of standards and practices. We then present the results of our analysis and offer observations and discussion of how those results are informing future planning for digital content management practices and processes.

A. Four Decades of Digital Collecting

The Library of Congress has four decades of experience acquiring, managing, and preserving digital collection content. In 1982 the Library's Optical Disk Pilot Project began capturing digitizing text and images of collections [1]. In 1989 those efforts grew into the American Memory pilot program, which digitized selected collection content for distribution and ultimately became the basis of the institution's first online collections [2].

These efforts have since grown to include a wide range of born-digital collecting programs. Of particular note, the Web Archiving Program launched in 2000, [3] copyright eDeposit for eSerials launched in 2009, [4] and the Cataloging-in-Publication program began to acquire eBooks in 2012 [5].

These programs, among many others, have resulted in continued growth of the Library's digital holdings. Building on the development of

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

DOI: 10.1145/nnnnnnn.nnnnnnn

these varied programs and initiatives, the Library of Congress established and published *Collecting Digital Content at the Library of Congress*, a five-year strategy and plan, which establishes targets in six strategic objectives for growing the digital collections [6]. In 2018, as a companion to its strategic plan, the Library of Congress established its first Digital Strategy which asserts the intention to continue to "exponentially grow" its digital collections and "ensure enduring access" to them [7].

As a result of the early and extensive efforts of the Library of Congress to build robust digitization and digital collecting programs, the institution has developed a digital collection that is, much like the physical collection, varied and expansive. The institution has simultaneously invested in a series of initiatives focused on ensuring enduring access to these parts of the library's collections.

B. Planning for Enduring Access to Content

In parallel to the growth and expansion of the Library of Congress digital collections, the institution has developed a series of resources to support planning for enduring access to these materials. Created and coordinated by a range of stakeholders in the Library, these include the development and maintenance of the Sustainability of Digital Formats website [8], the PREMIS Data Dictionary for Preservation Metadata [9], and the Library of Congress Recommended Formats Statement [10]. The institution also convenes stakeholders annually for the Designing Storage Architectures for Digital Collections summit. Along with these areas of work, the Library was a founding contributor to the National Digital Stewardship Alliance Levels of Digital Preservation [11].

All of these resources and initiatives have become critical components of international community approaches to digital preservation. Through the development of the Digital Collections Management Compendium (DCMC), the Library is currently working to synthesize and connect these interrelated resources.

C. Connecting Digital Content Theory & Practice

The Library of Congress is now working to develop a DCMC that summarizes current policies and internal guidance. The core concept for DCMC is to integrate and synthesize and formalize policy and guidance for practice, clarify roles and responsibilities, and clarify IT business needs. In this way, the effort provides a policy infrastructure that supports improvement of systems that acquire, preserve, and deliver digital content to users now and in the future.

The DCMC is being developed to connect high-level policy to the everyday work of digital content management practice. As a result, the Library's Digital Content Management section (DCM) is working to simultaneously integrate these various standards and planning resources into ongoing work to review the extent of existing digital collections and plan and implement processes and workflows to support future growth.

II. Identifying File Extensions in the Collections

In 2018, DCM was tasked with identifying and analyzing file formats in use in the Library's permanent digital collections. Ideally, analysis of file formats would involve direct analysis and characterization of the files for file signatures. However, it was not feasible to do so across all digital collection content with our current diverse digital library infrastructure. Instead, analysis was conducted against file extensions (e.g., .pdf, .mp3), which provide valuable information for understanding the likely formats of files. (We discuss the limitations of this approach below.)

Using our existing tools and data, DCM compiled and analyzed the file extensions of content inventoried in the Content Transfer System (CTS), an internally developed system that serves as the Library's primary inventory management system for digital collection content. CTS provides logs and inventory data for all content managed through the system across a range of distinct storage systems.

CTS is one of two approved inventory systems for managing digital collections content at the Library; the second is the Packard Campus Workflow Application (PCWA), which inventories digital collection content for the Motion Picture, Broadcasting and Recorded Sound (MBRS) Division. Currently, the collections managed by CTS comprise over 7 PB of digital content; the collections managed by PCWA comprise approximately 9 PB of content. Digital collections inventoried in PCWA are managed by MBRS and are not addressed in this analysis.

The results of DCM's analysis illustrate the extensive diversity of the Library's permanent digital collections. For the "long tail" of collection materials representing the diversity of file extensions, the Library works to ensure bit-level preservation, as reflected in the DCMC. CTS contains data that was used to report on file extensions of digital content in the collections, but this analysis also illustrates a series of potential next steps for improving inventory data for collection materials en masse. In particular, a

significant portion of files in the Library's digital collections are managed within compressed container files (ex. zip and tar), which mask the file extensions contained within them.

A. Scope of File Extension Data

DCM staff used Kibana, a data visualization platform, to create a custom dashboard to report extension data for CTS inventoried content. The Dashboard is not available to external users, but the development of it may be of general interest to external users.

The Dashboard uses the file extension index generated from CTS inventories to present data on file count and file size by extension. Dashboard users have the ability to filter file extension data by server system and custodial divisionsand can download various sets of data based on their categorization.

For both the Dashboard and the present report, we focused on a subset of the content inventoried in CTS that best represents a view of content in the permanent digital collection. This includes digital collection content under inventory control in long-term (tape) storage and access presentation (spinning disk) storage. Digital content in areas used for processing and ingest is excluded from the analysis.

This analysis includes all the files managed in these systems, including metadata files and other supporting files. All of this non-collection content and its respective file formats are under Library of Congress inventory control and are therefore content the institution is managing as part of the files in its permanent digital collection.

III. Analyzing File Extension Data

As of September 24, 2018, digital content under inventory control in the Library of Congress' primary inventory system (CTS) included approximately 681 million files represented in over 16,000 unique file extensions and occupying almost 8 petabytes of storage space,. It should be noted that these numbers include distinct copies of files in presentation and long-term storage. Due to aspects of the Library's digital infrastructure and the difficulty presented in de-duping inventories at the file level, an identical file stored in multiple locations or servers is counted multiple times for each discreet copy on disk.

We were also able to gain insights about the most common file extensions by file count and by file size. (The tables below list the top ten file extensions by count and size.) Further analysis of these results is presented in the next section, including discussion of the prevalence of image files by file count, the nature of compressed container files, and the unique case of files with no extension appearing to be the tenth most popular file across the Library's digital collections.

TABLE I
TOP 10 FILE EXTENSIONS BY COUNT

Rank	Extension	File Count	File Size
1	jp2	148,242,701	294.83 TB
2	tif	133,388,164	2,516.69 TB
3	jpg	113,952,865	36.40 TB
4	xml	70,117,369	41.98 TB
5	pdf	56,048,470	81.59 TB
6	txt	48,637,417	3.25 TB
7	gif	44,082,987	0.74 TB
8	gz	10,120,046	3,937.79 TB
9	i41	7,425,410	1.96 TB
10		5,091,599	3.16 TB

TABLE 2
TOP 10 FILE EXTENSIONS BY FILE SIZE

Rank	Extension	File Count	File Size
1	gz	10,120,046	3,937.79 TB
2	tif	133,388,164	2,516.69 TB
3	mxf	22,920	543.19 TB
4	jp2	148,242,701	294.83 TB
5	mpg	161,656	175.01 TB
6	wav	596,130	127.18 TB
7	mov	29,024	99.03 TB
8	pdf	56,048,470	81.59 TB
9	iso	36,325	63.74 TB
10	dv	5,006	54.48 TB

Analysis of file extension data suggests the following findings:

A) The majority of Library of Congress digital content—both by file count and by size—has file extensions associated with a relatively small number of file formats.

B) The remainder of file extensions reflect the Library of Congress's wide collecting scope as well as the scale and variation of the digital world.

C) The available set of tools for this analysis produced valuable information on Library of Congress digital collections, but an understanding of the current technical limitations suggests paths forward to improved analysis and monitoring.

A. Most Content has Well Known Extensions

By file size, about 80% of Library of Congress digital content is represented by two file extensions: .gz and .tif. Given the nature of current Library of Congress digital collections, it is not surprising that these two file extensions represent a significant majority by file size. The extension .gz is a compressed container format used, in this case, primarily for holding Web Archives content. This and other compressed container formats are discussed in more detail below. The extension .tif is an image format used extensively across many types of digitized content, including textual, photographic, and map content, as specified in the Recommended Formats Statement [10].

By file count, about 95% of Library of Congress digital content is represented by the following eight file extensions: .jp2, .tif, .jpg, .gif .xml, .txt, .pdf, and .gz. The image formats represent both master files and their corresponding derivative images, largely from digitized collection materials, reflecting the Library's digitization practices. As these files may correspond to a single page of a book or a single-image item such as a photograph, higher numbers are expected compared to, for example, born-digital ebooks, for which a single EPUB or PDF file usually represents a single book, a smaller but growing category of content at the Library. The textual formats may be collection content such as born-digital materials encoded in XML or plain text, but also include derivative content such as transcriptions, or metadata and supporting materials. Files with the extension .pdf may also represent born-digital master files or derivatives of digitized content.

Given the volume of Web Archives, the extension .gz also appears on this list of top file counts. The table below provides a break down of the extent to which these container files are associated with the Web Archives.

TABLE 3
TOP 10 FILE EXTENSIONS BY FILE COUNT

Extension	Total	Non Web Arch	Web Arch
gz	10,120,046	1,132,045	8,988,001
zip	410,092	409,878	214
rar	3,314	3,314	0
tar	3,026	3,015	11

TABLE 4
TOP 10 FILE EXTENSIONS BY FILE SIZE

Extension	Total	Non Web Arch	Web Arch
gz	3497.9 TB	560.7 TB	2987.7 TB
zip	15.7 TB	15.7 TB	0 TB
rar	0.2 TB	0.2 TB	0 TB
tar	1.9TB	1.9 TB	0 TB

The Web Archives are stored as WARCs, and then compressed into the .gz format. Display of the web archives content requires its own type of index. Those indexes allow for further analysis of the content of these .gz files. At the time of analysis in 2018, these web archive files contained 6.2 billion unique files [12].

With improvements to the Library's digital inventory systems, DCM will be able to provide more information about Library of Congress digital collections and bring more collection content under inventory control. This initial analysis suggests a large majority of current Library of Congress digital collection content is contained in a small number of file formats, which are widely supported and well understood for purposes of preservation and access. Large portions of the current collections are digitized surrogates of physical items, represented by the prevalence of image formats produced through digitization. As born-digital collecting expands in scope and content areas, we expect to see a greater variety of file formats in CTS.

B) Long Tail of File Extensions

Beyond the initial list of widely used file extensions exist a considerable long tail of extensions. Most of the 16,689 unique file extensions appear infrequently but provide a window into the diversity of the digital collections.

Across the Library of Congress digital collections, 2,761 file extensions are included

only once; most of these likely do not represent a unique file format, but instead represent a single file of an unknown format. In many cases, the "extension" part of the file name may have been used as a note to a user or a system, or a filename may incidentally contain a period, which is parsed in the data as representing an extension.

Furthermore, 14,064 file extensions, or 83% of all extensions, occur less than 100 times. It is very likely that some of these extensions represent collection material of a digital format for which further analysis is needed to determine more information. However, it is also likely that much of this content does not represent individual file formats, but instead are sub-components of dataset formats or are simply multiple copies of a file with the issue noted above: an unknown format where the "extension" field was possibly used as a note to a user or a system.

Finally, 3,810 files representing 272 file extensions are 0 byte files, meaning there is no content to the file except for a filename. These files may be purposefully empty to serve as a flag or identifier to a user or a system, or may expose corrupted files where the content was lost due to a system error.

The digital collections include content ranging from formats familiar to the average computer user and for which the preservation risks are well understood, to highly specialized and idiosyncratic formats. In some cases, as particularly seen in files created before the mid-1990s, file extensions were used as general purpose indicators or for sorting and may not have any relationship to file format at all.

In order to manage content associated with this long tail of file extensions, the DCMC provides guidance on the preservation and maintenance of original file formats. For some born-digital content, such as digital archival manuscript collections, the original content and file formats delivered to the Library may have artifactual qualities that need to be preserved for researchers. Such material will need to be maintained with its original context in addition to any potential migration or emulation for access.

C) Limitations Require System Improvements

Further understanding of the limitations of this analysis provides valuable information for improvements to digital inventory systems.

i. Limitations of Extension Information

The extensions data enabled some analysis and conclusions to be drawn about the Library's digital collections; however, file extensions remain the weakest form of file characterization. In the most striking example, files with no discernable, conventional extension—no characters following a period within the filename—are the 10th most popular "extension," representing 5,091,599 files and 3.161 TB of data. It is likely that most of this content is related to system functions, scripted operations, or datasets, but more advanced format analysis is required to determine if any of this content represents known file formats that should be managed as collection materials.

Another limitation is highlighted by the large subsets of extensions that can represent groupings of the same file format. In some cases these groupings are clearly defined; otherwise, they require further analysis to determine similarities. In this area, two major groupings emerged: datasets and system files.

Datasets: At least 20 different file extensions such as .i41, .i21, .i22, etc, are different extensions but effectively the same file format related to geospatial datasets. When these are separated out individually by extension they appear to be less prevalent, but combined they represent over 10 million files, or 1.5% of the total CTS inventoried content by file count. There are likely many other file extensions that are broken out as separate elements in this analysis, but in fact represent a single file format within a dataset.

System files: As with datasets, file formats used by systems in logging or tracking functions can result in misleading counts when analyzed by file extension. For example, .jdb files are stored as 2,388 different extensions, such as ".jdb,9999998", representing 154,411 total files; all of this content represents a single file format related to logging information for a certain type of Web Archives crawl action.

Library of Congress collections contain file extensions that include notes or references created by a user, in particular with digital content created decades ago, which is often donated to custodial divisions working with archival collections. These are valuable components of the provenance of the complete collection, but mask the functional file format of the content from current analysis.

ii. CTS Inventory Data Limitations

The CTS inventory data utilized for this analysis must be qualified for its significant limitations. For example, some portion the Library of Congress's digital collections are not inventoried by the CTS system and therefore not included in this analysis. (As previously mentioned, the MBRS content is managed in the

separate PCWA system.) Additionally, some portion of other digital collection content is currently in process to be brought under inventory control in CTS. While initial estimates indicate that a large majority of the long-term storage system is inventoried in CTS, DCM estimates that potentially as much as one third of the content (by file size) held in presentation server storage is not inventoried in CTS and therefore not represented in this analysis. DCM is thus establishing procedures for inventorying all content and maintaining the health of the inventory system.

In addition to the limitations created by uninventoried content, a potentially significant number of files are counted multiple times due to duplications and overlaps in inventory records. Duplication is expected for files stored in multiple server systems (long-term and presentation), but overlaps in inventory records result in the same discreet file on a server being counted twice. In order to improve the underlying data of CTS inventory records and improve further efforts to characterize the extent of digital content managed through this system, DCM has initiated an ongoing process to establish, improve, and refine use of the functions of the inventory system.

Given current system limitations, DCM is treating all content inventoried in CTS as "collection content." The CTS inventory system cannot currently filter the data into categories such as collection material or supporting files, metadata, or derivatives. CTS treats all files equally, since all are digital content under inventory control and managed in server systems that require monitoring. However, it may be helpful in the future to differentiate between born-digital and digitized collection content; masters from derivative, metadata, or supporting content; as this may determine preservation actions.

VI: Next Steps: Improve Inventory & Infrastructure

The results of this analysis offer substantive opportunities to expand work on building and managing digital collections at the Library of Congress. The creation of the internal Kibana Dashboard of file extensions provides an opportunity to observe both work to improve the inventory control and management of digital collections and further enhancements to digital content management tools and infrastructure.

A) Work to Improve Inventories

One of the primary results of this analysis has been to identify and prioritize work necessary to improve the underlying data used to track and monitor content in the digital collections.

We now have a baseline view into the state of the collection, and can plan projects to improve management and inventory control. DCM's future efforts to bring uninventoried collection content under inventory control will result in more trustworthy and complete reports and data on the entirety of the Library of Congress's digital collections. Furthermore, DCM is in the process of establishing regular data integrity checks of all content managed in CTS, which will ensure the inventory data remains in line with the current state of content. Additional remediation of metadata fields within CTS inventories will also improve the reportability of the data, such as assisting with filtering by custodial divisions and distinguishing master files from derivatives or supporting materials.

B) Potential Infrastructure Improvements

In addition to opportunities to improve the management of digital content, this analysis suggests a series of potential infrastructure improvements that could result in improved analysis of file formats in the digital collections.

1) Expansion of data available to reporting tools, including data on content in the other inventory system (PCWA), which would enable a wider variety of analyses.

2) Implementation of existing format characterization tools and validators (TIKA, JHOVE2) on a collection-level scale. Within existing systems, these tools can currently be run on individual resources, but to support planning and analysis, it is necessary to be able to be run these kinds of tools at scale against collection content.

3) Capability to generate and analyze data on contents of container files.

4) Ability to generate reporting across all instances of files, allowing for identification of duplicate files across systems for analysis and counting of unique files. Current deduplication is only possible at the larger inventory record level.

C) Anchoring Policy & Planning in Data

The results of this work serve as direct inputs for development of policy and planning resources for the institution. As we work to create and share an integrated DCMC of policies and practices, we are working to ensure that this resource bridges visions for how digital content management should work in theory and the realities of digital collection content in hand.

Based on this analysis, the DCMC clarifies the intention to maintain content in its original formats and offer levels of planned future support for digital content in alignment with information from the Sustainability of Digital Formats website and the Library of Congress Recommended Formats Statement. Additionally, as this data and analysis provides a view into all CTS inventoried content in the digital collections, it can be used to inform baseline practices for content management, help shape priorities for work on the Sustainability of Digital Formats, and help to inform future revisions to the Library's Recommended Formats Statement.

We opened this paper contextualizing our work in an understanding of the four decades of digital content management practice at the Library of Congress. A significant result of those four decades of work was making the case and then establishing the Digital Content Management unit. As we look to the next four decades and beyond, it will be increasingly essential to engage in the kind of analysis we have described in this report and put that analysis in dialog with both ongoing practices and the development of policy and planning resources.

REFERENCES

[1] M. Flynn and H. Zinkham, "The MARC Format and Electronic Reference Images: Experiences from the Library of Congress Prints and Photographs Division," *Visual Resources*, vol. 11, no. 1, pp. 47–70, Jan. 1995.

[2] C. R. Arms, "Historical collections for the National Digital Library: Lessons and challenges at the Library of Congress," *D-Lib Magazine April: Part*, vol. 1, 1996.

[3] A. Grotke, "Web Archiving at the Library of Congress," *Computers and Libraries*, vol. 31, no. 10, Dec. 2011.

[4] E. Delfino and J. Mandelbaum, "eDeposit for eSerials: Current Work and Plans at the Library of Congress," in *Journal Article Tag Suite Conference (JATS-Con) Proceedings*, 2011.

[5] T. Westervelt, "To Lead to Learning, Not to Madness: E-Books and E-Serials at the Library of Congress," *The Serials Librarian*, vol. 72, no. 1–4, pp. 122–127, May 2017.

[6] J. Puccio, "Collecting Digital Content at the Library of Congress," *The Signal*, 21-Mar-2017 https://blogs.loc.gov/thesignal/2017/03/collecting-digital-content-at-the-library-of-congress/.

[7] K. Zwaard, "Digital Strategy at the Library of Congress," presented at the Federal Depository Library Conference, Washington DC, Oct-2018.

[8] C. Arms and C. Fleischhauer, "Digital formats: Factors for sustainability, functionality, and quality," in *Archiving Conference*, 2005 pp. 222–227.

[9] PREMIS Data Dictionary, http://www.loc.gov/standards/premis

[10] T. Westervelt, "Library of Congress's Recommended Format Specifications," *Serials Review*, vol. 40, no. 3, pp. 217–218, Jul. 2014.M.

[11] Phillips, J. Bailey, A. Goethals, and T. Owens, "The NDSA Levels of Digital Preservation: An Explanation and Uses," *IS&T Archiving, Washington, USA*, 2013.

[12] G. Thomas and C. Dooley, "The Library of Congress Web Archives: Dipping a Toe in a Lake of Data," The Signal, 2019. http://blogs.loc.gov/thesignal/2019/01/the-library-of-congress-web-archives-dipping-a-toe-in-a-lake-of-data/